2314

EIN VERFAHREN ZUR ABSCHÄTZUNG DES UNTERSCHIEDES ZWISCHEN DEN LÖSUNGEN BENACHBARTER OPERATORGLEICHUNGEN

L. Wetterling

Hamburg 1961

Ein Verfahren zur Abschätzung des Unterschiedes zwischen den Lösungen benachbarter Operatorgleichungen

Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Hamburg

vorgelegt von
Wolfgang Wetterling
aus Insterburg

Hamburg
1961

Genehmigt von der

Mathematisch-Naturwissenschaftlichen Fakultät

der Universität Hamburg

auf Antrag von Professor Dr. Lothar Collatz

Hamburg, den 8. Februar 1961

Professor Dr. Erich Otremba

Dekan

R. UNIV. BIBLIOTHEEK LEIDEN

Inhaltsverzeichnis

	Seite
§ 1. Einleitung	5
§ 2. Einige Definitionen und ein Satz aus der Funktionalanalysis	8
§3. Problemstellung	10
§ 4. Ein bekanntes, nicht exaktes Abschätzungsverfahren	11
§ 5. Ein exaktes Abschätzungsverfahren	12
§ 6. Abschätzung bei Gleichungen (allgemeiner Fall)	15
§ 7. Abschätzung bei algebraischen Gleichungen	17
§ 8. Abschätzung bei linearen Gleichungssystemen	26
§ 9. Abschätzung bei Matrix-Eigenwertaufgaben	36
§ 10. Abschätzung bei linearen Integralgleichungen 1. Art	44
§ 11. Abschätzung bei gewöhnlichen Differentialgleichungen 1. Ordnung (Ersetzung einer nichtlinearen durch eine lineare Differentialgleichung)	49
§ 12. Beispiele	54
§ 13. ALGOL-Programme	68

§1. Einleitung

Bei numerischen Rechnungen kommt es häufig vor, daß eine gegebene Aufgabe nicht gelöst werden soll oder kann und daß statt dessen eine Ersatzaufgabe gelöst wird, die durch irgendwelche Abänderungen aus der gegebenen Aufgabe hervorgeht. Dieser Übergang kann beispielsweise notwendig sein, wenn die Ausgangswerte einer Rechnung mit Fehlern behaftet sind. Die gegebene Aufgabe ist dann die Rechnung mit den exakten, aber unbekannten Ausgangswerten, die Ersatzaufgabe ist die Rechnung mit bekannten, aber ungenauen Näherungen für die Ausgangswerte. Ein anderes Beispiel ist die Ersetzung einer vielleicht nur mit großem Aufwand zu lösenden nichtlinearen Aufgabe durch eine lineare Aufgabe, die leichter gelöst werden kann.

Wenn man eine Lösung der Ersatzaufgabe kennt, möchte man wissen, ob die gegebene Aufgabe eine Lösung besitzt, deren Abweichung von der Lösung der Ersatzaufgabe man an Hand der Abänderungen, die von der gegebenen Aufgabe zur Ersatzaufgabe führen, abschätzen kann. Es ist das Ziel dieser Arbeit, für Aufgaben, die sich als Operatorgleichungen (mit gewissen Eigenschaften) formulieren lassen, hinreichende Bedingungen für die Existenz einer solchen Lösung und Schranken für den Unterschied zwischen der Lösung der gegebenen Aufgabe und der Lösung der Ersatzaufgabe anzugeben. In der englischsprachigen Literatur werden solche Paare von Aufgaben - gegebene Aufgabe und durch Abänderungen entstehende Ersatzaufgabe - als „neighbour problems" bezeichnet [2]. In diesem Sinn ist der Ausdruck „benachbarte Operatorgleichungen" im Titel dieser Arbeit zu verstehen.

Die Herleitung der hinreichenden Bedingungen und der Abschätzung geschieht mit den Methoden der Funktionalanalysis. Die Begriffe, die dabei benötigt werden, sind in §2 definiert; außerdem ist dort ohne Beweis ein wichtiger Satz der Funktionalanalysis, der Fixpunktsatz für das allgemeine Iterationsverfahren in BANACHschen Räumen, angegeben. In §3 wird das zu behandelnde Problem, das hier schon angedeutet worden ist, genauer formuliert. In §4

wird ein Verfahren beschrieben, das für spezielle Aufgaben (algebraische Gleichungen, lineare Gleichungssysteme) in verschiedenen Lehrbüchern dargestellt ist und das eine allerdings nicht exakte Abschätzung für den Unterschied zwischen den Lösungen benachbarter Operatorgleichungen liefert. Bei der Herleitung dieses Verfahrens in den genannten Spezialfällen werden sogenannte „Produkte kleiner Größen" vernachlässigt. Eine dieser „kleinen Größen" ist aber immer der Unterschied der Lösungen der beiden Operatorgleichungen, also die Größe, die man abschätzen möchte, von der man zeigen möchte, daß sie klein ist.

Es ist daher wohl kein übertriebenes Streben nach Genauigkeit, wenn man jene Vernachlässigungen vermeiden möchte und nach einem exakten Abschätzungsverfahren sucht. Ein solches Verfahren wird in § 5 beschrieben. Es zeigt sich, daß nach Einführung eines Operators, der eine FRECHETsche Ableitung oder eine Näherung für diese sein kann, der Unterschied zwischen den Lösungen der benachbarten Operatorgleichungen selbst einer Operatorgleichung genügt, bei der der Fixpunktsatz für das Iterationsverfahren in BANACHschen Räumen angewendet werden kann. Falls dessen Voraussetzungen erfüllt sind, folgt die Existenz einer Lösung der gegebenen Aufgabe, deren Unterschied zu der Lösung der Ersatzaufgabe man abschätzen kann.

Die Herleitung dieser Abschätzungsmethode verläuft ähnlich wie die des NEWTONschen Iterationsverfahrens für allgemeine Operatorgleichungen. Bei der Anwendung auf spezielle Aufgaben sind ähnliche Überlegungen und Rechnungen wie bei diesem durchzuführen. Auch scheint die Anwendbarkeit auf jene Typen von Operatorgleichungen beschränkt zu sein, die sich mit dem NEWTONschen Iterationsverfahren behandeln lassen.

In den §§ 6 - 11 dieser Arbeit wird dann die in § 5 beschriebene allgemeine Abschätzungsmethode auf spezielle Operatorgleichungen angewendet. Dabei ergibt sich in § 10 eine Abschätzung für Integralgleichungen, die eine Verallgemeinerung einer bei KANTOROWITSCH-KRYLOW [6] elementar hergeleiteten Abschätzung ist, die nicht in den Bereich der in § 4 beschriebenen

unexakten Abschätzungen fällt, sondern eine exakte Abschätzung im Sinne von § 5 ist. § 12 bringt einige numerische Beispiele und § 13 in der algorithmischen Formelsprache ALGOL abgefaßte Programme zur Durchführung einiger der beschriebenen Abschätzungen auf elektronischen Rechenanlagen.

Über die Art, in der die Hilfsmittel der Funktionalanalysis in dieser Arbeit verwendet werden, ist noch zu bemerken, daß man gelegentlich mit allgemeineren als den BANACHschen Räumen und mit allgemeineren Fixpunktsätzen als dem hier angegebenen arbeitet und damit bei manchen Problemen allgemeinere und bessere Resultate erzielt. Der Versuch, auch bei dem hier behandelten Problem so zu verfahren, hat gezeigt, daß mit der Einführung von sogenannten pseudometrischen Räumen, bei denen zu je zwei Elementen ein Abstand definiert ist, der selbst Element einer halbgeordneten Menge ist [8] , mit wesentlich größerem Aufwand die gleichen Ergebnisse erzielt werden wie bei der Verwendung von BANACHschen Räumen, wenn man dabei in der Normdefinition Gewichtsfaktoren oder Gewichtsfunktionen benutzt. Auch der Versuch, mit einer Verallgemeinerung des Fixpunktsatzes durch Einführung sogenannter nichtlinearer Majoranten [9] zu arbeiten, hat keine günstigeren Ergebnisse geliefert.

Schließlich muß noch erwähnt werden, daß man das hier beschriebene Abschätzungsverfahren natürlich nur mit der Einschränkung ein exaktes Verfahren nennen kann, daß bei den zur Abschätzung notwendigen Rechnungen Abrundungsfehler auftreten, die dazu führen können, daß die berechneten Schranken nicht mehr exakt sind. In § 8 wird für den Spezialfall der linearen Gleichungssysteme gezeigt, wie der Einfluß der Abrundungsfehler verringert werden kann. Dort wird das allgemeine Abschätzungsverfahren von § 5 zunächst so angewendet, daß die Inverse der FRECHETschen Ableitung, in diesem Fall die zur Koeffizientenmatrix des Gleichungssystems inverse Matrix zur Abschätzung herangezogen wird, dann aber auch so, daß eine Näherung für die inverse Matrix benutzt wird. Diese Näherung kann die Matrix sein, die beim Prozeß der Matrixinversion unter dem Einfluß der Abrundungsfehler entsteht. So können die Abrundungsfehler

wenigstens zum Teil ausgeschaltet werden. Bei der Abschätzung muß man dafür eine zusätzliche Matrizenmultiplikation durchführen, die selbst wieder Anlaß zu Abrundungsfehlern gibt; diese sind aber im allgemeinen kleiner als die bei der Inversion auftretenden Abrundungsfehler.

§ 2. Einige Definitionen und ein Satz aus der Funktionalanalysis

Die in allen Lehrbüchern der Funktionalanalysis (z.B. [7]) zu findende Definition des BANACHschen Raumes soll hier nicht wiederholt werden. In § 3 werden zwei BANACHsche Räume R_1 und R_2 eingeführt. Die Bezeichnungen für die Normen in R_1 und R_2 müßten eigentlich verschieden sein. Da aber keine Verwechslungen zu befürchten sind, werden beide Normen in der gleichen Weise dargestellt: Die Norm eines Elementes x wird mit $\|x\|$ bezeichnet.

Es folgen nun einige Definitionen und ein wichtiger Satz, die später benutzt werden.

Definition: Die Menge aller Elemente x eines BANACHschen Raumes R, die der Ungleichung $\|x - x_0\| \leq r$ genügen, wobei x_0 ein festes Element von R und r eine positive reelle Zahl ist, heißt Kugel um x_0 mit dem Radius r.

Definition: Eine Umgebung eines Elementes $x \in R$ ist eine offene Teilmenge von R, die x enthält.

Definition: Ein Operator T ist eine Vorschrift, durch die jedem Element x aus einer Teilmenge D eines BANACHschen Raumes R_1 eindeutig ein Element Tx eines BANACHschen Raumes R_2 zugeordnet wird. D heißt Definitionsbereich von T. Die Menge W aller Tx mit $x \in D$ heißt Wertebereich von T.

Definition: Ein Operator T heißt linear, wenn für $x \in D$, $x' \in D$ stets $T(cx+c'x') = cTx + c'Tx'$ ist, wobei c und c' beliebige reelle bzw. komplexe Zahlen sind.

Definition: Ein Operator T, dessen Definitionsbereich D in einem BANACHschen Raum R_1 und dessen Wertebereich in einem BANACHschen Raum R_2 liegt, heißt FRECHET-differenzierbar, wenn es einen linearen Operator L,

ebenfalls mit Definitionsbereich in R_1 und Wertebereich in R_2, gibt, mit dem für $\|z\| \leq k$

$$\|T(u+z) - Tu - Lz\| \leq \|z\| \cdot \varepsilon(\|z\|)$$

ist, wobei k eine positive reelle Konstante und $\varepsilon(\lambda)$ eine nichtnegative reellwertige Funktion mit $\lim_{\lambda \to 0} \varepsilon(\lambda) = 0$ ist. Für den linearen Operator L schreibt man auch $T'_{(u)}$ und nennt ihn die FRECHETsche Ableitung von T an der Stelle u.

Fixpunktsatz für das Iterationsverfahren in BANACHschen Räumen:

H sei ein Operator, dessen Definitionsbereich D und Wertebereich W Teilmengen desselben BANACHschen Raumes R sind. z_o und z_1 seien zwei Elemente von R, für die $z_1 = Hz_o$ gilt. Wenn es eine positive Zahl $K < 1$ gibt, so daß

1. Die Kugel Ω aller Elemente $z \in R$ mit

   $\|z - z_1\| \leq \frac{K}{1-K} \|z_1 - z_o\|$ zu D gehört und

2. für je zwei Elemente z_I und z_{II} aus Ω die Ungleichung $\|Hz_I - Hz_{II}\| \leq K \cdot \|z_I - z_{II}\|$ gilt,

liegen alle Elemente der nach der Iterationsvorschrift $z_{\nu+1} = Hz_\nu$ $(\nu = 0,1,2,\ldots)$ gebildeten Folge $z_1, z_2, z_3, \ldots$ in Ω (und damit in D, die Iteration ist also unbeschränkt ausführbar), es gibt genau ein Element $z \in \Omega$, das der Gleichung $z = Hz$ genügt, und die Folge der z_ν konvergiert gegen dieses z.

Dieser bekannte Satz soll hier nicht bewiesen werden. Der bei COLLATZ [3] für eine etwas andere Formulierung des Satzes angegebene Beweis läßt sich fast wörtlich für die obige Formulierung übernehmen. Die in dem Satz auftretende Zahl K wird meist als LIPSCHITZ-Konstante bezeichnet.

§ 3. Problemstellung

S und T seien zwei Operatoren, deren Definitionsbereiche D_S und D_T Teilmengen eines BANACHschen Raumes R_1 sind. Die Wertebereiche W_S und W_T seien Teilmengen eines BANACHschen Raumes R_2. W_T enthalte das Nullelement Θ_2 von R_2, und es sei ein Element $u \in D_T$ bekannt, das durch T auf Θ_2 abgebildet wird, das also der Gleichung

$$Tu = \Theta_2 \tag{3.1}$$

genügt. u liege auch in D_S. T sei nun in irgendeiner Weise als Näherungsoperator für S gewählt. Dann erhebt sich die Frage, ob die Nullstelle u des Operators T eine Näherung für eine Nullstelle v des Operators S ist, ob nämlich in einer Kugel um u mit einem angebbaren Radius ein Element $v \in R_1$ existiert, das durch S auf Θ_2 abgebildet wird, das also der Gleichung

$$Sv = \Theta_2 \tag{3.2}$$

genügt. S sei ein Operator, bei dem zu einem Element $x \in R_1$ das Bildelement Sx nicht bekannt ist, sondern nur eine Teilmenge von R_2, in der Sx liegt. Insbesondere sei von Su nur bekannt, daß es in einer Kugel um Θ_2 liegt, daß also $\|Su\|$ nicht größer als eine Schranke σ ist.

Das Ziel dieser Arbeit ist, hinreichende Bedingungen für die Existenz einer Lösung v von (3.2) und, wenn diese Bedingungen erfüllt sind, eine Schranke für $\|v - u\|$ anzugeben. Dazu ist zunächst vorauszusetzen, daß ein Operator Q mit einem Definitionsbereich $D_Q \subset R_1$ und einem Wertebereich $W_Q \subset R_2$ existiert, der folgende Eigenschaften hat:

1. D_Q enthält eine Umgebung U von Θ_1.
2. Q bildet Θ_1 auf Θ_2 ab: $Q\Theta_1 = \Theta_2$.
3. Die Zuordnung der $z \in U$ zu ihren Bildelementen Qz ist umkehrbar eindeutig, d.h. der inverse Operator Q^{-1} existiert.

Einige Voraussetzungen über U folgen später. Q ist so zu wählen, daß Qz für $z \in U$ eine Näherung für T(u+z) ist, wobei u die Lösung von (3.1) ist. Q kann beispielsweise die FRECHETsche Ableitung $T'_{(u)}$ sein, falls diese und der zu ihr inverse Operator $(T'_{(u)})^{-1}$ in einer passenden Umgebung von Θ_1 bzw. Θ_2 exi-

stieren. In fast allen Anwendungsbeispielen der §§ 6 - 11 wird Q die FRECHETsche Ableitung $T'_{(u)}$ sein.

Es soll nun $v = u + z$ gesetzt und (3.2) in

$$Qz - Qz + T(u+z) - T(u+z) + S(u+z) = 0_2$$

und weiter in

$$z = Q^{-1}[Qz - T(u+z) + T(u+z) - S(u+z)] \qquad (3.3)$$

umgeformt werden. Wenn v der Gleichung (3.2) genügt und $z = v - u$ in U liegt, erfüllt z die Gleichung (3.3). Wenn andererseits z in U liegt und der Gleichung (3.3) genügt, erfüllt $v = u + z$ die Gleichung (3.2). Aus der somit zu (3.2) äquivalenten Gleichung (3.3) sollen nun in § 4 und § 5 Folgerungen gezogen werden.

§ 4. Ein bekanntes, nicht exaktes Abschätzungsverfahren

Bei BODEWIG [2] werden für lineare Gleichungssysteme (S. 141) und bei WILLERS [10] für algebraische Gleichungen (S. 279) Abschätzungsverfahren angegeben, mit denen man keine exakten Schranken für $\|v - u\|$ erhält, sondern nur eine Aussage, wie groß $\|v - u\|$ etwa werden kann. Dabei wird die für die einzelnen Spezialfälle aufgestellte Gleichung (3.3) dadurch vereinfacht, daß „Produkte kleiner Größen" vernachlässigt werden. In der hier angegebenen Form der Problemstellung können jene Verfahren in folgender Weise formuliert werden:

Als Operator Q wird die FRECHETsche Ableitung $T'_{(u)}$ verwendet. Die Differenz $T'_{(u)} - T(u+z)$ wird - unter der Annahme, daß z „klein" ist - als vernachlässigbar klein angesehen und weggelassen (bei linearen Problemen, also bei linearen Gleichungssystemen, verschwindet sie tatsächlich), und bei der „kleinen" Differenz $T(u+z) - S(u+z)$ wird die kleine Verschiebung des Arguments von $u + z$ nach u vorgenommen, so daß (3.3) in

$$z \approx (T'_{(u)})^{-1}(Tu - Su) = -(T'_{(u)})^{-1}Su$$

übergeht. Wenn man nun mit Hilfe der Schranke σ für $\|Su\|$ eine Schranke τ für $\|(T'_{(u)})^{-1}Su\|$ angeben kann, weiß man nicht, ob τ auch eine Schranke für $\|z\| = \|v - u\|$ ist. Man

vertraut aber darauf, daß $\|v-u\|$ nicht viel größer als τ werden kann, wenn die an (3.3) vorgenommenen Vereinfachungen dadurch gerechtfertigt sind, daß die vernachlässigten Größen tatsächlich klein sind.

Im folgenden § 5 soll nun ein exaktes Verfahren angegeben werden, das diese Vernachlässigungen vermeidet und die gewünschten hinreichenden Bedingungen für die Existenz einer Lösung v von (3.2) und eine Schranke für $\|v-u\|$ liefert.

§ 5. Ein exaktes Abschätzungsverfahren

Die Gleichung (3.3) kann als Operatorgleichung

$$z = Hz \tag{5.1}$$

mit $Hz = Q^{-1}[Qz - S(u+z)] = z - Q^{-1}S(u+z)$ geschrieben werden. Mit $z_0 = \Theta_1$ beginnend soll nach der Vorschrift $z_{\nu+1} = Hz_\nu$ ($\nu=0,1,2,\ldots$) die Folge $z_0, z_1, z_2, \ldots$ gebildet werden. Insbesondere ist

$$z_1 = Hz_0 = -Q^{-1}Su.$$

Wenn die Voraussetzungen des Fixpunktsatzes für Iterationsverfahren in BANACHschen Räumen erfüllt sind, konvergiert die Folge der z_ν gegen ein Element z, das Lösung der Gleichung (5.1) ist. Die Voraussetzungen des Fixpunktsatzes besagen, daß für je zwei Elemente z_I und z_{II} einer Kugel Ω

$$\|Hz_I - Hz_{II}\| \leq K \cdot \|z_I - z_{II}\| \tag{5.2}$$

mit einer positiven LIPSCHITZ-Konstante $K<1$ sein muß. Ω ist dabei die Kugel um z_1, die alle Elemente $\zeta \in R_1$ mit

$\|\zeta - z_1\| \leq \frac{K}{1-K}\|z_1 - z_0\|$, also mit

$$\|\zeta - z_1\| \leq \frac{K}{1-K}\|z_1\| \tag{5.3}$$

enthält. Nun ist nach § 3 Su und damit z_1 nicht bekannt, sondern nur eine Schranke σ für $\|Su\|$. Es sei möglich, aus dieser Schranke σ eine Schranke τ für $\|z_1\| = \|Q^{-1}Su\|$ zu bestimmen. Da nach der Dreiecksungleichung $\|\zeta\| \leq \|\zeta - z_1\| + \|z_1\|$ ist, folgt aus (5.3) die Ungleichung

$$\|\zeta\| \leq \frac{K}{1-K}\|z_1\| + \|z_1\| = \frac{1}{1-K}\|z_1\|$$

und hieraus

$$\|\zeta\| \leq \frac{1}{1-K}\tau. \tag{5.4}$$

Die Kugel Ω ist also in der durch (5.4) beschriebenen Kugel Ω' um Θ_1 enthalten. Wenn man nun ein $K<1$ so bestimmen kann, daß sogar für je zwei Elemente z_I und z_{II} von Ω' die Ungleichung (5.2) erfüllt ist, sind die Voraussetzungen des Fixpunktsatzes erfüllt. Es folgt dann, daß in Ω eine Lösung z von (5.1) liegt. Da nicht z_1 selbst, sondern nur eine Schranke τ für $\|z_1\|$ bekannt ist, ist wiederum nur die Aussage möglich, daß z in Ω' liegt, daß also eine Lösung z von (5.1) mit

$\|z\| \leq \frac{1}{1-K}\tau$ existiert.

Um den Fixpunktsatz in dieser Weise überhaupt anwenden zu können, muß man wissen, ob H in Ω' definiert ist. Hierzu muß $\mathfrak{D}_S$, der Definitionsbereich von S, alle Elemente u+z von R_1 mit $z \in \Omega'$ enthalten, und der Bereich, in dem Q^{-1} definiert ist, also der Bereich, auf den die in § 3 angegebene Umgebung U von O_1 durch Q umkehrbar eindeutig abgebildet wird, muß die Menge aller S(u+z) mit $z \in \Omega'$ enthalten. Die letzte Bedingung ist sicher erfüllt, wenn, wie bei allen Anwendungen in den §§ 6 - 11, Q ein linearer Operator ist. Wenn nämlich ein linearer Operator Q eine Umgebung U von $\Theta_1 \in R_1$ umkehrbar eindeutig auf eine Umgebung von $\Theta_2 \in R_2$ abbildet, bildet er den ganzen BANACHschen Raum R_1 umkehrbar eindeutig auf R_2 ab. Andernfalls wäre $Qx = Qx'$ für $x \neq x'$ ($x \in R_1$, $x' \in R_1$), und für eine genügend kleine positive reelle Zahl η lägen die (verschiedenen) Elemente ηx und $\eta x'$ in U und würden auf dasselbe Element $\eta Qx = \eta Qx'$ abgebildet. Das ist ein Widerspruch.

Wenn die Voraussetzungen des Fixpunktsatzes nachgeprüft sind, weiß man, daß es ein Element $z \in R_1$ mit $z = Hz$ gibt. Daraus folgt, daß $Q^{-1}S(u+z) = \Theta_1$ ist. Man möchte aber wissen, ob $v = u + z$ der Gleichung (3.2) genügt, ob also $S(u+z) = \Theta_2$ ist. Nach der zweiten Voraussetzung in § 3 über den Operator Q ist tatsächlich

$$\Theta_2 = Q\Theta_1 = QQ^{-1}S(u+z) = S(u+z).$$

Es gibt also eine Lösung v von $Sv = \Theta_2$ mit $\|v-u\| \leq \frac{\tau}{1-K}$.

Es soll noch gezeigt werden, daß die Gleichung $z = Hz$ in Ω' höchstens eine Lösung besitzt, wenn die Voraussetzungen des Fixpunktsatzes erfüllt sind. Gäbe es zwei verschiedene Elemen-

te z und z' in Ω' mit $z = Hz$ und $z' = Hz'$, so wäre $\|Hz - Hz'\| = \|z - z'\|$, entgegen der Voraussetzung, daß $\|Hz - Hz'\| \leq K\|z - z'\|$ mit $K < 1$ ist.

Wenn $Q = T'_{(u)}$ ist, unterscheidet sich die hier aufgestellte Schranke für $\|z\| = \|v - u\|$ von der in § 4 angegebenen Schranke um den Faktor $\frac{1}{1-K}$, der größer als 1 ist, wenn $0 < K < 1$ ist. Man erhält also gegenüber dem Verfahren von § 4 eine gröbere, dafür aber exakte Abschätzung.

Der Operator T tritt hier gar nicht mehr auf, ebenso Q selbst, sondern nur S und der inverse Operator Q^{-1}. T und Q wurden nur zur Formulierung des zu lösenden Problems in § 3 benutzt. Daher sollen auch bei der folgenden Zusammenfassung des Abschätzungsverfahrens in einem Satz nur die Operatoren S und Q^{-1}, wofür P gesetzt wird, verwendet werden.

Satz: S sei ein Operator, der eine Teilmenge D_S eines BANACHschen Raumes R_1 auf eine Teilmenge W_S eines BANACHschen Raumes R_2 abbildet. u sei ein Element aus D_S. Es sei möglich, einen Operator P, der eine Umgebung des Nullelements Θ_2 von R_2 auf eine Umgebung des Nullelements Θ_1 von R_1, Θ_2 selbst auf Θ_1 und sonst kein Element auf Θ_1 abbildet, so zu bestimmen, daß die folgenden Voraussetzungen erfüllt sind:

1. $-PSu$ existiert, und es ist $\|PSu\| \leq \tau$.
2. Es gibt eine Zahl K mit $0 < K < 1$, so daß
   a) $Hz = z - PS(u+z)$ für $\|z\| \leq \frac{\tau}{1-K}$ existiert,
   b) für je zwei Elemente z_I, z_{II} von R_1 mit $\|z_I\| \leq \frac{\tau}{1-K}$, $\|z_{II}\| \leq \frac{\tau}{1-K}$ die Ungleichung
   $$\|Hz_I - Hz_{II}\| \leq K\|z_I - z_{II}\|$$
   erfüllt ist.

Dann besitzt die Gleichung $Sv = \Theta_2$ genau eine Lösung v mit

$$\|v - u\| \leq \frac{\tau}{1-K}.$$

§ 6. Abschätzung bei Gleichungen (allgemeiner Fall)

Zunächst soll das Abschätzungsverfahren von § 5 auf Gleichungen mit einer Unbekannten angewendet werden. $R_1 = R_2 = R$ sei der BANACHsche Raum der reellen Zahlen x mit der Norm $\|x\| = |x|$. Die Operatorgleichungen (3.1) und (3.2) seien zwei Gleichungen

$$s(v) = 0, \quad t(u) = 0$$

zur Bestimmung der Unbekannten v bzw. u. Die Lösung u der zweiten Gleichung sei bekannt. t(x) sei bei u differenzierbar, und es sei $t'(u) \neq 0$. s(x) besitze in einer Umgebung von u eine stetige erste Ableitung. Als Operator Q soll die FRECHETsche Ableitung von T an der Stelle u gewählt werden. Da $T(u+z) = t(u+z) = t(u) + t'(u)\cdot z + z\cdot\varepsilon(z)$ mit $\lim \varepsilon(z) = 0$ ist, wird $T'_{(u)}z = t'(u)\cdot z$ und $(T'_{(u)})^{-1}h = \frac{h}{t'(u)} \cdot \overset{z\to 0}{\exists}\ (T'_{(u)})^{-1}h$ existiert also für alle h. Für den Operator H erhält man

$$Hz = z - \frac{s(u+z)}{t'(u)} ,$$

und der erste Iterationsschritt ergibt mit $z_0 = 0$

$$z_1 = Hz_0 = -\frac{s(u)}{t'(u)} .$$

Zur Berechnung der LIPSCHITZ-Konstante K bildet man

$$Hz_I - Hz_{II} = z_I - z_{II} - \frac{s(u+z_I) - s(u+z_{II})}{t'(u)} .$$

Nach dem Mittelwertsatz der Differentialrechnung wird

$$Hz_I - Hz_{II} = \frac{t'(u) - s'(u+\zeta)}{t'(u)}(z_I - z_{II})$$

für ein ζ zwischen z_I und z_{II}. σ sei eine Schranke für $|s(u)|$. Wenn es eine positive Konstante $K < 1$ gibt, so daß

$$\left|\frac{t'(u) - s'(u+\zeta)}{t'(u)}\right| \leq K$$

in dem Bereich $|\zeta| \leq \frac{\tau}{1-K}$ mit $\tau = \frac{\sigma}{|t'(u)|}$ ist, folgt nach § 5, daß es eine Lösung v von $s(v) = 0$ mit

$$|v - u| \leq \frac{\tau}{1-K}$$

gibt.

Eine Abschätzung nach § 4 würde so erfolgen: Man würde für s(x) den Ansatz $s(x) = t(x) + r(x)$ machen, würde in

$$t(u+z) + r(u+z) \approx t(u) + zt'(u) + r(u) + zr'(u) = 0 .$$

das Produkt $zr'(u)$ vernachlässigen und $z = -\frac{t(u)+r(u)}{t'(u)} = -\frac{s(u)}{t'(u)}$ und daraus $|z| \leq \frac{\sigma}{|t'(u)|}$ erhalten. Beide Abschätzungen sind in Abb. 1 dargestellt.

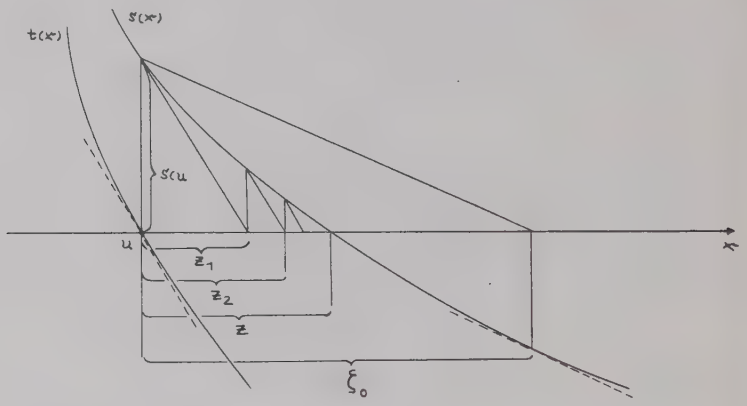

Abb. 1

$z_1 = -\frac{s(u)}{t'(u)}$ ist die eine Kathete des rechtwinkligen Dreiecks, dessen andere Kathete $s(u)$ ist und dessen Hypotenuse parallel zur Tangente an die Kurve $t(x)$ bei $x = u$ ist. Ebensolche Dreieckskonstruktionen liefern auf Grund der Gleichungen $z_{\nu+1} - z_\nu = -\frac{s(u+z_\nu)}{t'(u)}$ die Strecken $z_2, z_3, \ldots$ Die Strecke $\frac{1}{1-K}\left|\frac{s(u)}{t'(u)}\right|$ kann ebenfalls zeichnerisch ermittelt werden. Bei dem gezeichneten Beispiel soll nur der Bereich $x > u$ untersucht werden. Dort ist $|s'(x)| < |t'(u)|$ und daher

$$\left|\frac{t'(u) - s'(u+\xi)}{t'(u)}\right| = 1 - \frac{s'(u+\xi)}{t'(u)}$$

für positives ξ. K soll so bestimmt werden, daß im Intervall

$$0 \leq \xi \leq \frac{1}{1-K}\left|\frac{s(u)}{t'(u)}\right|$$

$$\operatorname{Max}\left[1 - \frac{s'(u+\xi)}{t'(u)}\right] = K$$

ist. Da $\frac{s'(u+\xi)}{t'(u)}$ für wachsendes ξ monoton abnimmt, wird das Maximum am rechten Intervallende angenommen, d.h. für $\xi_o = \frac{1}{1-K}\left|\frac{s(u)}{t'(u)}\right|$ wird $K = 1 - \frac{s'(u+\xi_o)}{t'(u)}$ und $\frac{1}{1-K} = \frac{t'(u)}{s'(u+\xi_o)}$.

Hieraus ergibt sich $\xi_0 = \left|\frac{s(u)}{s'(u+\xi_0)}\right|$ als Bestimmungsgleichung für ξ_0. ξ_0 ist also die eine Kathete des rechtwinkligen Dreiecks, dessen andere Kathete $s(u)$ ist und dessen Hypotenuse parallel zur Tangente an die Kurve $s(x)$ bei $x = u + \xi_0$ ist. Die durch die Abschätzung nach § 4 erhaltene Schranke für z ist z_1, z wird also unterschätzt; durch die Abschätzung nach § 5 erhält man die Schranke $\xi_0 > z$.

§ 7. Abschätzung bei algebraischen Gleichungen

Die Ergebnisse von § 6 sollen nun auf algebraische Gleichungen angewendet werden. Es sei

$$t(x) = \sum_{k=0}^{n} a_k x^k \ , \quad s(x) = \sum_{k=0}^{n} (a_k + c_k) x^k \ .$$

Von den c_k sei nur bekannt, daß ihre Beträge durch γ_k beschränkt sind: $|c_k| \leq \gamma_k$ $(k=0,1,\ldots,n)$. Eine einfache Nullstelle u von $t(x)$ sei bekannt, es sei also

$$t(u) = \sum a_k u^k = 0 \ , \quad t'(u) = \sum k a_k u^{k-1} \neq 0.$$

Dann wird $|s(u)| = |\sum c_k u^k| \leq \sum \gamma_k |u|^k = \sigma$. Ferner ist nach dem Mittelwertsatz der Differentialrechnung

$$\left|\frac{t'(u) - s'(u+\xi)}{t'(u)}\right| = \left|\frac{\sum k a_k u^{k-1} - \sum k(a_k + c_k)(u+\xi)^{k-1}}{\sum k a_k u^{k-1}}\right|$$

$$= \left|\frac{\sum k c_k (u+\xi)^{k-1} + \xi \cdot \sum k(k-1) a_k (u+\xi')^{k-2}}{\sum k a_k u^{k-1}}\right| ,$$

wobei ξ' eine Zahl zwischen 0 und ξ ist. Daher wird

$$\left|\frac{t'(u) - s'(u+\xi)}{t'(u)}\right| \leq \varphi(|\xi|) =$$

$$= \frac{\sum k \gamma_k (|u|+|\xi|)^{k-1} + |\xi| \cdot \sum k(k-1)|a_k| (|u|+|\xi|)^{k-2}}{|\sum k a_k u^{k-1}|} \ .$$

Die LIPSCHITZ-Konstante K soll so bestimmt werden, daß

$\varphi(|\xi|) \leq K$ für $|\xi| \leq \frac{1}{1-K} \frac{\sum \gamma_k |u|^k}{|\sum k a_k u^{k-1}|} = \xi^*$ ist. In diesem Bereich nimmt $\varphi(|\xi|)$ sein Maximum für $|\xi| = \xi^*$ an. Wenn man für

K dieses Maximum, also $\varphi(\xi^*)$ wählt, wird

$$\xi^* = \frac{1}{1-\varphi(\xi^*)} \frac{\sum \gamma_k |u|^k}{|\sum k a_k u^{k-1}|} = f(\xi^*) \qquad (7.1)$$

eine Gleichung zur Bestimmung von ξ^*. Man kann ξ^* iterativ berechnen, indem man mit $\xi_0 = 0$ oder $\xi_0 = \frac{\sum \gamma_k |u|^k}{|\sum k a_k u^{k-1}|}$ beginnt und das Iterationsverfahren $\xi_{\nu+1} = f(\xi_\nu)$ $(\nu=0,1,2,\ldots)$ durchführt. Wenn die LIPSCHITZ-Konstante $K<1$ ist, konvergiert die Folge der ξ_ν gegen eine Lösung ξ^* der Gleichung (7.1). Das folgt daraus, daß für $K<1$ die Folge der ξ_ν monoton wachsend und beschränkt ist. Für $0 \leqq \xi' < \xi'' \leqq \xi^*$ ist dann nämlich $0 < f(\xi') < f(\xi'') \leqq \xi^*$, und es ist $\xi_1 > \xi_0$. Wenn andererseits die Folge der ξ_ν gegen eine Zahl ξ^* konvergiert, ist diese Lösung der Gleichung (7.1). $\varphi(\xi^*)$ ist dann als LIPSCHITZ-Konstante K verwendbar. Wenn diese kleiner als 1 ist, existiert eine Nullstelle v von s(x) mit

$$|v-u| \leqq \xi^* = \frac{1}{1-K} \frac{\sum \gamma_k |u|^k}{|\sum k a_k u^{k-1}|} . \qquad (7.2)$$

Das gefundene Ergebnis soll in einem Satz zusammengefaßt werden.

Satz: Das Polynom $t(x) = \sum_{k=0}^{n} a_k x^k$ besitze die einfache Nullstelle u. s(x) sei ein Polynom, dessen Koeffizienten sich von den a_k höchstens um γ_k unterscheiden. Wenn die Folge $\xi_0, \xi_1, \xi_2, \ldots$, die, beginnend mit $\xi_0 = 0$ oder

$$\xi_0 = \frac{\sum_{k=0}^{n} \gamma_k |u|^k}{\left|\sum_{k=1}^{n} k a_k u^{k-1}\right|}$$

, nach der Vorschrift

$$\varphi(\xi_\nu) = \frac{\sum_{k=1}^{n} k\gamma_k (|u|+\xi_\nu)^{k-1} + \xi_\nu \sum_{k=2}^{n} k(k-1)|a_k| (|u|+\xi_\nu)^{k-2}}{\left|\sum_{k=1}^{n} k a_k u^{k-1}\right|} ,$$

$$\zeta_{\nu+1} = \frac{1}{1-\varphi(\zeta_\nu)} \frac{\sum_{k=0}^{n} \gamma_k |u|^k}{\left|\sum_{k=1}^{n} k a_k u^{k-1}\right|} \qquad (\nu=0,1,2,\ldots)$$

gebildet wird, gegen eine positive Zahl ζ^* mit $\varphi(\zeta^*) < 1$ konvergiert, wobei stets $\varphi(\zeta_\nu) < 1$ bleibt, besitzt das Polynom s(x) eine Nullstelle v mit

$$|v-u| \leq \frac{1}{1-K} \frac{\sum_{k=0}^{n} \gamma_k |u|^k}{\left|\sum_{k=1}^{n} k a_k u^{k-1}\right|} .$$

Dabei ist $K = \varphi(\zeta^*)$. Wenn für irgendein ν $\varphi(\zeta_\nu) \geq 1$ wird, ist keine derartige Aussage über eine Nullstelle von s(x) möglich.

Die Abschätzung (7.2) ohne den Faktor $\frac{1}{1-K}$, die nach dem in § 4 beschriebenen Prinzip gewonnen wird, findet sich bei WILLERS [10], S. 279.

Man kann für algebraische Gleichungen beim Vorliegen eines Paares von konjugiert komplexen Nullstellen eine Abschätzung herleiten, die auf das doppelzeilige HORNER-Schema zur Bestimmung von quadratischen Faktoren eines Polynoms zurückgeht ([5], [13]). Diese Methode ist auch bei Doppelwurzeln von t(x) und bei nahe an einer Nullstelle von t'(x) liegenden reellen Wurzelpaaren von t(x) anwendbar, wo die Abschätzung (7.2) wegen des Verschwindens oder der Kleinheit von t'(x) nicht mehr durchführbar ist. Wenn t(x) einen quadratischen Faktor $x^2 + ax + b$ enthält und die Voraussetzungen für die Abschätzung erfüllt sind, wird sich ergeben, daß s(x) einen Faktor $x^2 + a'x + b'$ enthält, wobei man für $|a' - a|$ und $|b' - b|$ Schranken angeben kann. Man kann also etwa im Falle einer Doppelwurzel von t(x) angeben, wie weit sich diese beim Übergang zu s(x) in ein Paar von konjugiert komplexen oder reellen Nullstellen auflösen kann.

Es sei wieder

$$t(x) = \sum_{k=1}^{n} a_k x^k , \quad s(x) = t(x) + r(x) = t(x) + \sum_{k=1}^{n} c_k x^k .$$

Wenn man $t(x)$ durch ein quadratisches Polynom $x^2 + ax + b$ dividiert, ergibt sich

$$t(x) = (x^2+ax+b)t_1(x) + xA_t(a,b) + B_t(a,b) \ , \qquad (7.3)$$

wobei $t_1(x)$ ein Polynom (n-2). Grades in x ist. $A_t(a,b)$ und $B_t(a,b)$ sind Polynome in den zwei Variablen a und b von höchstens (n-1). Grade (vgl. HOUSEHOLDER [5], S. 139-140). Entsprechende Polynome $A_s(a,b)$, $B_s(a,b)$, $A_r(a,b)$, $B_r(a,b)$ erhält man, wenn man $s(x)$ oder $r(x)$ durch $x^2 + ax + b$ dividiert. Wenn $x^2 + ax + b$ in $t(x)$ als Faktor enthalten ist, verschwindet $xA_t(a,b) + B_t(a,b)$ identisch in x. Daraus folgt

$$A_t(a,b) = B_t(a,b) = 0 \ . \qquad (7.4)$$

Umgekehrt folgt aus diesen Gleichungen, daß $t(x)$ den Faktor $x^2 + ax + b$ enthält. Die Bestimmung des quadratischen Faktors $x^2 + ax + b$ von $t(x)$ ist damit auf die Lösung des Systems (7.4) von zwei algebraischen Gleichungen für a und b zurückgeführt.

Zur Herleitung einer Abschätzung nach § 5 wählt man als BANACHschen Raum $R_1 = R_2 = R$ (vgl. §3) den Raum der Paare reeller Zahlen $\binom{a}{b}$ mit der Norm

$$\left\| \binom{a}{b} \right\| = \mathrm{Max}(\alpha|a|, \beta|b|) \ .$$

α und β sind dabei noch festzulegende positive reelle Zahlen. Den Operator T definiert man durch

$$T\binom{a}{b} = \begin{pmatrix} A_t(a,b) \\ B_t(a,b) \end{pmatrix} \ ,$$

ebenso definiert man S durch

$$S\binom{a}{b} = \begin{pmatrix} A_s(a,b) \\ B_s(a,b) \end{pmatrix} \ .$$

Zunächst soll $T'_{\binom{a}{b}}$ bestimmt werden. Man findet

$$T'_{\binom{a}{b}}\binom{c}{d} = \begin{pmatrix} \frac{\partial A_t}{\partial a} & \frac{\partial A_t}{\partial b} \\ \frac{\partial B_t}{\partial a} & \frac{\partial B_t}{\partial b} \end{pmatrix} \binom{c}{d}$$

(Produkt einer Matrix mit einem Vektor). Wenn man in (7.3) $t_1(x)$ wiederum durch $x^2 + ax + b$ dividiert, erhält man

$$t(x) = (x^2+ax+b)^2 t_2(x) + (x^2+ax+b)\left[xC_t(a,b) + D_t(a,b)\right] + \\ + xA_t(a,b) + B_t(a,b) \; . \qquad (7.5)$$

Auch $C_t(a,b)$ und $D_t(a,b)$ sind Polynome der zwei Variablen a und b. Durch partielle Differentiation nach a und b erhält man aus (7.5) - wenn man berücksichtigt, daß wohl $t_2(x)$, nicht aber t(x) von a und b abhängt - die folgenden Gleichungen:

$$2(x^2+ax+b)xt_2(x) + (x^2+ax+b)^2 \frac{\partial t_2}{\partial a} + x^2 C_t + xD_t + \\ + (x^2+ax+b)\left(x\frac{\partial C_t}{\partial a} + \frac{\partial D_t}{\partial a}\right) + x\frac{\partial A_t}{\partial a} + \frac{\partial B_t}{\partial a} = 0$$

$$2(x^2+ax+b)t_2(x) + (x^2+ax+b)^2 \frac{\partial t_2}{\partial b} + xC_t + D_t + \\ + (x^2+ax+b)\left(x\frac{\partial C_t}{\partial b} + \frac{\partial D_t}{\partial b}\right) + x\frac{\partial A_t}{\partial b} + \frac{\partial B_t}{\partial b} = 0 \; .$$

x_o sei eine Nullstelle von $x^2 + ax + b$. Dann folgt durch Einsetzen von $x_o^2 = -ax_o - b$

$$\left.\begin{aligned} x_o\left(-aC_t + D_t + \frac{\partial A_t}{\partial a}\right) + \left(-bC_t + \frac{\partial B_t}{\partial a}\right) &= 0 \\ x_o\left(C_t + \frac{\partial A_t}{\partial b}\right) + \left(D_t + \frac{\partial B_t}{\partial b}\right) &= 0 \; . \end{aligned}\right\} \qquad (7.6)$$

Wenn es zwei verschiedene Nullstellen x_o von $x^2 + ax + b$ gibt (dann ist $a^2 \neq 4b$), wenn die Gleichungen (7.6) also für zwei verschiedene x_o erfüllt sind, müssen die vier in Klammern stehenden Ausdrücke verschwinden, es wird dann

$$\frac{\partial A_t}{\partial a} = aC_t - D_t \qquad \frac{\partial A_t}{\partial b} = -C_t$$

$$\frac{\partial B_t}{\partial a} = bC_t \qquad \frac{\partial B_t}{\partial b} = -D_t \; .$$

Als Gleichungen zwischen Polynomen in a und b sind diese Gleichungen aus Stetigkeitsgründen auch für $a^2 = 4b$ gültig, also auch dann, wenn $x^2 + ax + b$ das Quadrat eines Polynoms ersten Grades ist und nur eine Nullstelle besitzt.

Der inverse Operator $\left(T'_{\binom{a}{b}}\right)^{-1}$ existiert, wenn die Determinante

$$\frac{\partial A_t}{\partial a}\frac{\partial B_t}{\partial b} - \frac{\partial A_t}{\partial b}\frac{\partial B_t}{\partial a} = D_t^2 - aC_tD_t + bC_t^2$$

von Null verschieden ist. Sollte die Determinante verschwin-

den, so ist entweder $C_t = 0$ und daher auch $D_t = 0$, oder $-\frac{D_t}{C_t}$ ist eine Nullstelle von $x^2 + ax + b$. Wenn man voraussetzt, daß $T\binom{a}{b} = 0$ ist, enthält im ersten Fall $t(x)$ den Faktor $(x^2+ax+b)^2$; im zweiten Fall ist $xC_t + D_t$ ein Faktor von $x^2 + ax + b$. Falls $x^2 + ax + b$ eine Doppelwurzel besitzt, ist diese dann nach (7.5) eine mindestens dreifache Nullstelle von $t(x)$. Falls $x^2 + ax + b$ zwei verschiedene Wurzeln besitzt, ist eine von ihnen eine Doppelwurzel von $t(x)$. Wenn diese Fälle ausgeschieden werden, existiert der inverse Operator $\left(T'\binom{a}{b}\right)^{-1}$.

Man muß nun zunächst eine Schranke für $\left\|S\binom{a}{b}\right\|$ berechnen. Aus $s(x) = t(x) + r(x)$ folgt

$$A_s(a,b) = A_t(a,b) + A_r(a,b) ,$$
$$B_s(a,b) = B_t(a,b) + B_r(a,b) ,$$

da der Rest, der bei der Division einer Summe von Polynomen entsteht, der Summe der bei der Division der einzelnen Polynome entstehenden Reste gleich ist. Wenn $x^2 + ax + b$ in $t(x)$ als Faktor enthalten ist und daher $A_t = B_t = 0$ ist, wird

$$A_s(a,b) = A_r(a,b) , \quad B_s(a,b) = B_r(a,b) .$$

Wenn für die Koeffizienten von $r(x)$ Schranken bekannt sind, wenn nämlich $|c_k| \leq \gamma_k$ ist, erhält man Schranken $\bar{A}_r$ für $|A_r(a,b)|$ und $\bar{B}_r$ für $|B_r(a,b)|$ und damit eine Schranke für $\left\|S\binom{a}{b}\right\|$, wenn man den Divisionsalgorithmus (doppelzeiliges HORNER-Schema) mit den Polynomen $\sum \gamma_k x^k$ und $x^2 - |a|x - |b|$ durchführt. Wenn man nun die Beträge der 4 Elemente der Matrix $\left(T'\binom{a}{b}\right)^{-1}$ mit T_{11}, T_{12}, T_{21}, T_{22} bezeichnet, wird

$$\left\|\left(T'\binom{a}{b}\right)^{-1} S\binom{a}{b}\right\| \leq \mathrm{Max}\left[\alpha(T_{11}\bar{A}_r + T_{12}\bar{B}_r),\ \beta(T_{21}\bar{A}_r + T_{22}\bar{B}_r)\right] = \tau .$$

Damit hat man die nach § 5 notwendige Schranke τ.

Nun können die in der Normdefinition auftretenden Konstanten α und β so bestimmt werden, daß sich eine günstige Abschätzung ergibt. Wenn eine Abschätzung überhaupt möglich ist, wird nach § 5

$$\|v - u\| = \left\|\binom{c}{d}\right\| \leq \frac{1}{1-K}\tau ,$$

und daher wird

$$|c| \leq \frac{1}{1-K}\frac{\tau}{\alpha} \quad , \quad |d| \leq \frac{1}{1-K}\frac{\tau}{\beta} \, .$$

α und β sollen so gewählt werden, daß $\frac{\tau}{\alpha}$ und $\frac{\tau}{\beta}$ möglichst klein werden. Das ist recht einfach; wollte man dagegen bei der Bestimmung möglichst günstiger α und β auch noch den Faktor $\frac{1}{1-K}$ berücksichtigen, so würde man viel kompliziertere Vorschriften erhalten. Es ist

$$\frac{\tau}{\alpha} = \mathrm{Max}\left[(T_{11}\bar{A}_r+T_{12}\bar{B}_r), \frac{\beta}{\alpha}(T_{21}\bar{A}_r+T_{22}\bar{B}_r)\right] \, ,$$

$$\frac{\tau}{\beta} = \mathrm{Max}\left[\frac{\alpha}{\beta}(T_{11}\bar{A}_r+T_{12}\bar{B}_r), (T_{21}\bar{A}_r+T_{22}\bar{B}_r)\right] \, .$$

Die Maxima können nicht kleiner werden als die Ausdrücke ohne die Faktoren $\frac{\beta}{\alpha}$ und $\frac{\alpha}{\beta}$. Sie werden gleich diesen Ausdrücken, wenn man

$$\alpha = T_{21}\bar{A}_r+T_{22}\bar{B}_r \, , \quad \beta = T_{11}\bar{A}_r+T_{12}\bar{B}_r$$

setzt. Dann wird

$$|c| \leq \frac{1}{1-K}(T_{11}\bar{A}_r+T_{12}\bar{B}_r) \, , \quad |d| \leq \frac{1}{1-K}(T_{21}\bar{A}_r+T_{22}\bar{B}_r) \, .$$

Wenn $T_{11}\bar{A}_r+T_{12}\bar{B}_r = 0$ oder $T_{21}\bar{A}_r+T_{22}\bar{B}_r = 0$ ist, kann man α und β nicht auf diese Weise bestimmen. Das sind aber triviale Ausnahmefälle, die hier ausgeschlossen werden sollen.

Schließlich ist noch die LIPSCHITZ-Konstante K zu bestimmen. Mit $Q = T'_{\binom{a}{b}}$ ist (vgl. § 5)

$$H\binom{c}{d} = \binom{c}{d} - \left(T'_{\binom{a}{b}}\right)^{-1} S\binom{a+c}{b+d} \, .$$

Man erhält

$$H\binom{c_I}{d_I} - H\binom{c_{II}}{d_{II}} = \binom{c_I-c_{II}}{d_I-d_{II}} - \left(T'_{\binom{a}{b}}\right)^{-1}\binom{A_s(a+c_I,b+d_I)-A_s(a+c_{II},b+d_{II})}{B_s(a+c_I,b+d_I)-B_s(a+c_{II},b+d_{II})} \, .$$

Nun ist nach dem Mittelwertsatz der Differentialrechnung

$$A_s(a+c_I),b+d_I)-A_s(a+c_{II},b+d_{II}) =$$

$$= (c_I-c_{II})\frac{\partial A_s(a+c',b+d')}{\partial a} + (d_I-d_{II})\frac{\partial A_s(a+c',b+d')}{\partial b}$$

für ein c' zwischen c_I und c_{II} und ein d' zwischen d_I und d_{II}.

Ferner ist

$$\frac{\partial A_s(a+c',b+d')}{\partial a} = \frac{\partial A_t(a,b)}{\partial a} + c'\frac{\partial^2 A_t(a+\tilde{c},b+\tilde{d})}{\partial a^2} + \\ + \frac{\partial^2 A_t(a+\tilde{c},b+\tilde{d})}{\partial a \partial b} + \frac{\partial A_r(a+c',b+d')}{\partial a}$$

für ein $\tilde{c}$ zwischen 0 und c' und ein $\tilde{d}$ zwischen 0 und d'. Aus diesen und aus den für die anderen auftretenden Ausdrücke entsprechend gebildeten Gleichungen folgt

$$H\begin{pmatrix} c_I \\ d_I \end{pmatrix} - H\begin{pmatrix} c_{II} \\ d_{II} \end{pmatrix} = \begin{pmatrix} c_I - c_{II} \\ d_I - d_{II} \end{pmatrix} - \begin{pmatrix} \frac{\partial A_t}{\partial a} & \frac{\partial A_t}{\partial b} \\ \frac{\partial B_t}{\partial a} & \frac{\partial B_t}{\partial b} \end{pmatrix}^{-1} \cdot$$

$$\cdot \left\{ \begin{pmatrix} \frac{\partial A_t}{\partial a} & \frac{\partial A_t}{\partial b} \\ \frac{\partial B_t}{\partial a} & \frac{\partial B_t}{\partial b} \end{pmatrix} + \underbrace{\begin{pmatrix} c'\frac{\partial^2 \tilde{A}_t}{\partial a^2} + d'\frac{\partial^2 \tilde{A}_t}{\partial a \partial b} + \frac{\partial \tilde{A}_r}{\partial a} & c'\frac{\partial^2 \tilde{A}_t}{\partial a \partial b} + d'\frac{\partial^2 \tilde{A}_t}{\partial b^2} + \frac{\partial \tilde{A}_r}{\partial b} \\ c''\frac{\partial^2 \tilde{B}_t}{\partial a^2} + d''\frac{\partial^2 \tilde{B}_t}{\partial a \partial b} + \frac{\partial \tilde{B}_r}{\partial a} & c''\frac{\partial^2 \tilde{B}_t}{\partial a \partial b} + d''\frac{\partial^2 \tilde{B}_t}{\partial b^2} + \frac{\partial \tilde{B}_r}{\partial b} \end{pmatrix}}_{= M} \right\} \begin{pmatrix} c_I - c_{II} \\ d_I - d_{II} \end{pmatrix}$$

$$= -\left(T'_{\binom{a}{b}}\right)^{-1} \cdot M \begin{pmatrix} c_I - c_{II} \\ d_I - d_{II} \end{pmatrix} . \qquad (7.7)$$

Dabei ist die Ableitung an einer Zwischenstelle zu nehmen, wenn dies durch eine Tilde angedeutet ist, z.B.

$$\frac{\partial^2 \tilde{A}}{\partial a^2} = \frac{\partial^2 A(a+\tilde{c},b+\tilde{d})}{\partial a^2} .$$

Für die Beträge der Elemente der Matrix M kann man Schranken angeben, wenn die Argumente der auftretenden partiellen Ableitungen in Intervallen $[a-\zeta,a+\zeta]$, $[b-\eta,b+\eta]$ liegen und $|c'|$ und $|c''|$ nicht größer als ζ, $|d'|$ und $|d''|$ nicht größer als η sind. Diese Schranken gewinnt man, indem man auf $\sum |a_k| x^k$ und $\sum \gamma_k x^k$ das doppelzeilige HORNER-Schema für die Division durch $x^2 - (|a|+\zeta)x - (|b|+\eta)$ anwendet (vgl. ZURMÜHL [13], S. 51 ff.). Die Schranken sollen mit $M_{\mu\nu}(\zeta,\eta)$ (μ, ν = 1, 2) bezeichnet werden. Nach (7.7) erhält man eine in den obigen Intervallen gültige LIPSCHITZ-Konstante

$$K(\xi,\eta) = \mathrm{Max}\Big\{\big[T_{11}M_{11}(\xi,\eta)+T_{12}M_{21}(\xi,\eta)\big] + \frac{\alpha}{\beta}\big[T_{11}M_{12}(\xi,\eta)+T_{12}M_{22}(\xi,\eta)\big] ,$$
$$\frac{\beta}{\alpha}\big[T_{21}M_{11}(\xi,\eta)+T_{22}M_{21}(\xi,\eta)\big]+\big[T_{21}M_{12}(\xi,\eta)+T_{22}M_{22}(\xi,\eta)\big]\Big\} .$$

Wie bei einfachen Nullstellen sind die Zahlen ξ^* und η^*, die die Intervalle angeben, in denen der Operator H durch $K = K(\xi^*,\eta^*)$ LIPSCHITZ-beschränkt sein muß, damit die Abschätzung nach § 5 möglich ist, durch die Gleichungen

$$\xi^* = \frac{1}{1-K(\xi^*,\eta^*)}\,(T_{11}\bar{A}_r+T_{12}\bar{B}_r) = f_1(\xi^*,\eta^*) ,$$
$$\eta^* = \frac{1}{1-K(\xi^*,\eta^*)}\,(T_{21}\bar{A}_r+T_{22}\bar{B}_r) = f_2(\xi^*,\eta^*)$$

bestimmt. Man kann diese ebenfalls durch Iteration lösen, indem man von $\xi_o = \eta_o = 0$ oder $\xi_o = T_{11}\bar{A}_r+T_{12}\bar{B}_r$, $\eta_o = T_{21}\bar{A}_r+T_{22}\bar{B}_r$ ausgeht und die Folgen ξ_o, ξ_1, ξ_2,... und η_o, η_1, η_2,... nach der Vorschrift

$$\xi_{\nu+1} = f_1(\xi_\nu,\eta_\nu) \qquad (\nu=0,1,2,\ldots)$$
$$\eta_{\nu+1} = f_2(\xi_\nu,\eta_\nu)$$

berechnet. Diese Folgen konvergieren, falls $K<1$ ist, denn sie sind dann monoton wachsend und beschränkt. Das folgt daraus, daß die $M_{\mu\nu}(\xi,\eta)$ und daher auch $K(\xi,\eta)$ monoton wachsende Funktionen von ξ und η sind und daß $\xi_1 > \xi_o$, $\eta_1 > \eta_o$ ist. Wenn andererseits die Folgen der ξ_ν und η_ν gegen positive Zahlen ξ^* und η^* mit $K(\xi^*,\eta^*) = K < 1$ konvergieren, ist K als LIPSCHITZ-Konstante verwendbar, und die Existenz eines quadratischen Faktors $x^2+a'x+b'$ von s(x) mit

$$|a'-a| \leqq \xi^* \quad , \quad |b'-b| \leqq \eta^* \tag{7.8}$$

ist nachgewiesen. Man hat dann hieraus noch zu folgern, in welchem Bereich der reellen Achse oder der komplexen Zahlenebene Nullstellen

$$\nu = -\frac{a'}{2} \pm \sqrt{\left(\frac{a'}{2}\right)^2 - b'}$$

von s(x) liegen können.

Abschließend ist zu bemerken, daß man ebenso, wie man das doppelzeilige HORNER-Schema verallgemeinern kann zu einem Schema für die Division durch Polynome von höherem als zweitem Grade, die hier beschriebene Abschätzung auf den Fall verallgemeinern kann, daß t(x) einen Faktor höheren Grades enthält und nach einem ebensolchen von s(x) gesucht wird.

§ 8. Abschätzung bei linearen Gleichungssystemen

$R_1 = R_2 = R$ sei der n-dimensionale reelle Vektorraum. Im folgenden werden verschiedene Normdefinitionen verwendet, und zwar

$$\|x\| = \operatorname{Max}_i(|x_i|) , \tag{8.1}$$

als Verallgemeinerung hiervon

$$\|x\| = \operatorname{Max}_i(\alpha_i|x_i|) , \tag{8.2}$$

wobei die α_i positive reelle Zahlen sind, die später so gewählt werden, daß sich eine möglichst günstige Abschätzung ergibt, ferner

$$\|x\| = \sum_{i=1}^{n} |x_i| \tag{8.3}$$

und

$$\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2} \quad . \tag{8.4}$$

Die sich bei den verschiedenen Normdefinitionen ergebenden Abschätzungen bei linearen Gleichungssystemen sollen miteinander verglichen werden. Die Operatorgleichung $Tu = \theta_2$ laute

$$Au - r = 0 . \tag{8.5}$$

$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ - & - & - & - \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$ sei eine n-reihige, quadratische, nichtsinguläre Matrix. $u = \begin{pmatrix} u_1 \\ \cdots \\ u_n \end{pmatrix}$ und $r = \begin{pmatrix} r_1 \\ \cdots \\ r_n \end{pmatrix}$ seien Vektoren aus R. Die Lösung u von (8.5) sei bekannt. Die Operatorgleichung $Sv = \theta_2$ laute

$$(A+C)v - (r+s) = 0 . \tag{8.6}$$

C sei eine n-reihige quadratische Matrix, von deren Elementen c_{ik} nur bekannt ist, daß ihre Beträge durch γ_{ik} beschränkt sind: $|c_{ik}| \leq \gamma_{ik}$. s sei ein Vektor, von dessen Komponenten s_i ebenfalls nur bekannt ist, daß ihre Beträge durch σ_i beschränkt sind: $|s_i| \leq \sigma_i$. Durch die Anwendung des allgemeinen Abschätzungsverfahrens von § 5 findet man hinreichende Bedingungen dafür, daß es eine Lösung v von (8.6) gibt, deren Abweichung von u man abschätzen kann, bei der man nämlich Schranken für $\|v - u\|$ und $|v_i - u_i|$ angeben kann. Als Operator Q soll zunächst die FRECHETsche Ableitung $T'_{(u)}$ verwendet werden.

Es ist $T(u+z) = A(u+z) - r = (Au-r) + Az$ und daher $T'_{(u)}z = T'z = Az$ (die FRECHETsche Ableitung hängt hier nicht von u ab). Da A nichtsingulär sein soll, existiert der inverse Operator $(T')^{-1} = A^{-1}$. Für den Operator H (vgl. § 5) erhält man dann

$$Hz = z - A^{-1}\left[(A+C)(u+z) - (r+s)\right] .$$

Da u der Gleichung (8.5) genügt, wird $Hz = z - A^{-1}\left[Az + C(u+z) - s\right]$. Hieraus folgt

$$Hz = - A^{-1}\left[C(u+z) - s\right] . \tag{8.7}$$

Für $z_0 = 0$ wird $z_1 = Hz_0 = - A^{-1}\left[Cu - s\right]$. Für zwei Vektoren z_I, $z_{II} \in R$ wird

$$Hz_I - Hz_{II} = - A^{-1}C(z_I - z_{II}).$$

Die Elemente von A^{-1} sollen mit b_{ik} bezeichnet werden. Wenn man die Normdefinition (8.1) verwendet, erhält man

$$\begin{aligned} \|Hz_I - Hz_{II}\| &= \max_i \left(\Big| \sum_{j,k} b_{ij}c_{jk}(z_{I_k} - z_{II_k}) \Big| \right) \\ &\leq \max_i \left(\sum_{j,k} |b_{ij}|\,\gamma_{jk}\,|z_{I_k} - z_{II_k}| \right) \\ &\leq \max_i \left(\sum_{j,k} |b_{ij}|\,\gamma_{jk} \right) \|z_I - z_{II}\| \\ &= K_1 \|z_I - z_{II}\| \end{aligned}$$

mit $K_1 = \max_i \left(\sum_{j,k} |b_{ij}|\,\gamma_{jk} \right)$. Dies gilt für beliebige Vektoren z_I, $z_{II} \in R$. Daher existiert, falls $K_1 < 1$ ist, eine Lösung v von (8.6) mit

$$\begin{aligned} |v_l - u_l| \leq \max_i |v_i - u_i| = \|v - u\| \leq \\ \leq \frac{\tau}{1 - K_1} = \frac{\max_i \left(\sum_j |b_{ij}| \left(\sum_k \gamma_{jk}|u_k| + \sigma_j \right) \right)}{1 - K_1} \end{aligned} \tag{8.8}$$

$$(l=1,2,\ldots,n).$$

Bei Verwendung der Normdefinition (8.2) ergibt sich

$$\begin{aligned} \|Hz_I - Hz_{II}\| &= \max_i \left(\alpha_i \Big| \sum_{j,k} b_{ij}c_{jk}(z_{I_k} - z_{II_k}) \Big| \right) \\ &\leq \max_i \left(\alpha_i \sum_{j,k} |b_{ij}|\,\gamma_{jk}\frac{1}{\alpha_k} \right) \max_k \left(\alpha_k |z_{I_k} - z_{II_k}| \right) \\ &= K_2 \|z_I - z_{II}\| \end{aligned}$$

mit $K_2 = \max_i \left(\alpha_i \sum_{j,k} |b_{ij}|\,\gamma_{jk}\frac{1}{\alpha_k} \right)$. Falls $K_2 < 1$ ist, gibt es eine

Lösung v von (8.6) mit

$$\|v-u\| \leq \frac{\tau}{1-K_2} = \frac{\max_i\left\{\alpha_i \sum_j |b_{ij}|\left(\sum_k \gamma_{jk}|u_k| + \sigma_j\right)\right\}}{1-K_2} .$$

Hieraus folgt

$$|v_l-u_l| \leq \frac{1}{\alpha_l}\|v-u\|$$
$$\leq \frac{1}{1-K_2}\max_i\left\{\frac{\alpha_i}{\alpha_l}\sum_j |b_{ij}|\left(\sum_k \gamma_{jk}|u_k| + \sigma_j\right)\right\} \qquad (8.9)$$
$$(l=1,2,\ldots,n).$$

Wie in § 7 die Konstanten α und β, so sollen hier die Konstanten α_i so gewählt werden, daß für $l = 1,2,\ldots,n$

$$M_l = \max_i\left\{\frac{\alpha_i}{\alpha_l}\sum_j |b_{ij}|\left(\sum_k \gamma_{jk}|u_k| + \sigma_j\right)\right\}$$

möglichst klein wird. Für $i = l$ ist der Faktor $\frac{\alpha_i}{\alpha_l} = \frac{\alpha_l}{\alpha_l} = 1$. Daher ist, wie man die α_i auch wählt,

$$M_l \geq \sum_j |b_{lj}|\left(\sum_k \gamma_{jk}|u_k| + \sigma_j\right) .$$

Wenn man

$$\alpha_i = \frac{1}{\sum_j |b_{ij}|\left(\sum_k \gamma_{jk}|u_k| + \sigma_j\right)} \qquad (i=1,2,\ldots,n) \quad (8.10)$$

setzt, wird

$$M_l = \sum_j |b_{lj}|\left(\sum_k \gamma_{jk}|u_k| + \sigma_j\right) \qquad (l=1,2,\ldots,n) .$$

Man macht alle M_l gleichzeitig so klein wie möglich, wenn man die α_i nach (8.10) wählt. Es ist also nicht nötig, zur Minimisierung verschiedener M_l auch verschiedene Werte für die α_i zu nehmen. (8.9) wird jetzt zu

$$|v_l-u_l| \leq \frac{1}{1-K_2}\left\{\sum_j |b_{lj}|\left(\sum_k \gamma_{jk}|u_k| + \sigma_j\right)\right\} \quad (l=1,2,\ldots,n) \qquad (8.11)$$

Es kann sein, daß für einen oder mehrere Indizes i der Nenner in (8.10) Null wird. Dieser Fall muß hier ausgeschlossen werden und kann z. B. nach (8.8) behandelt werden.

Mit der Normdefinition (8.3) wird

$$\|Hz_r - Hz_\pi\| = \sum_i \left|\sum_{j,k} b_{ij}c_{jk}(z_{r_k} - z_{\pi_k})\right|$$
$$\leq \sum_{i,j,k} |b_{ij}|\cdot\gamma_{jk}\cdot|z_{r_k} - z_{\pi_k}| \leq$$

$$\leq \operatorname{Max}_k\Big(\sum_{i,j} |b_{ij}|\gamma_{jk}\Big) \sum_k |z_{I_k} - z_{II_k}|$$

$$= K_3 \, \|z_I - z_{II}\|$$

mit $K_3 = \operatorname{Max}_k\Big(\sum_{i,j} |b_{ij}|\gamma_{jk}\Big)$. Wenn $K_3 < 1$ ist, existiert eine Lösung v von (8.6) mit

$$\sum_i |v_i - u_i| = \|v-u\| \leq \frac{\tau}{1-K_3} = \frac{\sum_{i,j} |b_{ij}|\big(\sum_k \gamma_{jk}|u_k| + \sigma_j\big)}{1-K_3} \,. \qquad (8.12)$$

Verwendet man schließlich die Normdefinition (8.4), so wird

$$\|Hz_I - Hz_{II}\| = \sqrt{\sum_i \Big[\sum_{j,k} b_{ij} c_{jk} (z_{I_k} - z_{II_k})\Big]^2} \,.$$

Nach der SCHWARZschen Ungleichung folgt hieraus

$$\|Hz_I - Hz_{II}\| \leq \sqrt{\sum_i \Big[\sum_k \Big(\sum_j b_{ij} c_{jk}\Big)^2\Big] \cdot \Big[\sum_k (z_{I_k} - z_{II_k})^2\Big]}$$

$$\leq \sqrt{\sum_{i,k} \Big(\sum_j |b_{ij}|\gamma_{jk}\Big)^2} \sqrt{\sum_k (z_{I_k} - z_{II_k})^2}$$

$$= K_4 \, \|z_I - z_{II}\|$$

mit $K_4 = \sqrt{\sum_{i,k} \Big(\sum_j |b_{ij}|\gamma_{jk}\Big)^2}$. Wenn $K_4 < 1$ ist, gibt es eine Lösung v von (8.6) mit

$$\sqrt{\sum_i (v_i - u_i)^2} = \|v-u\| \leq \frac{\tau}{1-K_4} = \frac{\sqrt{\sum_i \Big[\sum_j |b_{ij}|\big(\sum_k \gamma_{jk}|u_k| + \sigma_j\big)\Big]^2}}{1-K_4} \,. \qquad (8.13)$$

Wenn alle γ_{jk} durch eine obere Schranke γ und alle σ_j durch eine obere Schranke σ ersetzt werden, können die Abschätzungsformeln (8.8), (8.11), (8.12), (8.13) vereinfacht werden (dabei wird $\sum_j |b_{ij}| = \beta_i$ gesetzt) zu

$$|v_l - u_l| \leq \frac{\operatorname{Max}_i(\beta_i)\big(\gamma \sum_k |u_k| + \sigma\big)}{1 - n\gamma \operatorname{Max}_i(\beta_i)} \,, \qquad (8.8^*)$$

$$|v_l - u_l| \leq \frac{\beta_l\big(\gamma \sum_k |u_k| + \sigma\big)}{1 - \operatorname{Max}_i \dfrac{\beta_i \gamma \sum_k \beta_k \big(\gamma \sum_m |u_m| + \sigma\big)}{\beta_i \big(\gamma \sum_m |u_m| + \sigma\big)}} =$$

$$= \frac{\beta_l(\gamma \sum_k |u_k| + \sigma)}{1 - \gamma \sum_k \beta_k} , \qquad (8.11^*)$$

$$\sum_i |v_i - u_i| \leq \frac{\sum_i \beta_i (\gamma \sum_k |u_k| + \sigma)}{1 - \gamma \sum_k \beta_k} , \qquad (8.12^*)$$

$$\sqrt{\sum_i (v_i - u_i)^2} \leq \frac{\sqrt{\sum_i \beta_i^2}\,(\gamma \sum_k |u_k| + \sigma)}{1 - \gamma \sqrt{n \sum_i \beta_i^2}} . \qquad (8.13^*)$$

(8.12*) erhält man auch direkt aus (8.11*), indem man in (8.11*) die Summen der rechten und der linken Seiten bildet. Die Ungleichung, die für alle Summanden gilt, gilt auch für die Summen.

Ein Vergleich der Abschätzungen (8.8) und (8.11) zeigt, daß die LIPSCHITZ-Konstante K_2 größer oder kleiner als die LIPSCHITZ-Konstante K_1 (oder ihr gleich) sein kann. Wenn alle $\gamma_{jk} = \gamma$ und alle $\sigma_j = \sigma$ sind, sieht man aus (8.8*) und (8.11*), daß $K_2 = \gamma \sum \beta_k < \gamma \cdot n \operatorname{Max}_i(\beta_i) = K_1$ sein kann. Um zu zeigen, daß auch $K_2 > K_1$ sein kann, wählt man als Matrix A die Einheitsmatrix E. Dann ist auch $A^{-1} = E$ und

$$K_1 = \operatorname{Max}_i(\sum_k \gamma_{ik}) , \quad K_2 = \operatorname{Max}_i(\sum_k \gamma_{ik} \frac{\alpha_i}{\alpha_k}) , \quad \alpha_k = \frac{1}{\sum_m \gamma_{km} |u_m| + \sigma_k} .$$

Wenn man für ein i_o, für das $K_1 = \sum_k \gamma_{i_o k}$ ist, die σ_k so wählt, daß $\alpha_k < \alpha_{i_o}$ für $k \neq i_o$ ist, kann

$$K_2 \geq \sum_k \gamma_{i_o k} \frac{\alpha_{i_o}}{\alpha_k} > \sum_k \gamma_{i_o k} = K_1$$

werden. Das in (8.8) auftretende $\operatorname{Max}_i \{ \sum_j |b_{ij}| (\sum_k \gamma_{jk} |u_k| + \sigma_j)$ ist nicht kleiner als der in (8.11) auftretende Ausdruck $\sum_j |b_{lj}| (\sum_k \gamma_{jk} |u_k| + \sigma_j)$. Wenn man von den Faktoren $\frac{1}{1 - K_1}$ und $\frac{1}{1 - K_2}$ absieht, liefert also (8.11) für gewisse l im allgemeinen kleinere Schranken für $|v_l - u_l|$ als (8.8). Die Abschätzungen (8.12) und (8.13) sind zur Berechnung von Schranken

für $|v_1-u_1|$ weniger geeignet als (8.8) und (8.11), da man dabei die Ungleichungen $|v_1-u_1| \leqq \sum |v_i-u_i|$ oder $|v_1-u_1| \leqq \sqrt{\sum (v_i-u_i)^2}$ verwenden müßte, die meistens zu einer sehr groben Abschätzung von $|v_1-u_1|$ führen.

In dem folgenden Satz ist das Ergebnis zusammengefaßt, das in der zur Abschätzung der einzelnen Differenzen $|v_1 - u_1|$ besonders günstig erscheinenden Ungleichung (8.11) enthalten ist.

Satz: $u_1, u_2, \ldots, u_n$ sei Lösung des linearen Gleichungssystems

$$\sum_{j=1}^{n} a_{ij} u_j = r_i \qquad (i=1,2,\ldots,n)$$

mit nichtsingulärer Koeffizientenmatrix A. C sei eine n-reihige quadratische Matrix und s ein Vektor, von deren Elementen c_{ij} und s_i nur bekannt ist, daß $|c_{ij}| \leqq \gamma_{ij}$ und $|s_i| \leqq \sigma_i$ ist mit nichtnegativen reellen Zahlen γ_{ij} und σ_i. b_{ij} seien die Elemente der inversen Matrix A^{-1}. Wenn mit

$$\alpha_i = \frac{1}{\sum_j |b_{ij}| \left(\sum_k \gamma_{jk} |u_k| + \sigma_j \right)}$$

die LIPSCHITZ-Konstante

$$K = \max_i \left(\sum_{j,k} |b_{ij}| \gamma_{jk} \frac{\alpha_i}{\alpha_k} \right) < 1$$

ist, besitzt das lineare Gleichungssystem

$$\sum_{j=1}^{n} (a_{ij}+c_{ij}) v_j = r_i + s_i \quad (i=1,2,\ldots,n)$$

eine Lösung $v_1, v_2, \ldots, v_n$ mit

$$|v_1-u_1| \leqq \frac{\sum_j |b_{1j}| \left(\sum_k \gamma_{jk} |u_k| + \sigma_j \right)}{1-K} \qquad (1=1,2,\ldots,n) .$$

Es soll nun diskutiert werden, welcher Rechenaufwand zur Ermittlung von Schranken für $\|v-u\|$ bzw. $|v_1-u_1|$ erforderlich ist. Dabei sind zwei Punkte zu berücksichtigen:

1. Die Berechnung von u und von A^{-1}.

   Als Maß für den Rechenaufwand bei der Lösung eines linearen Gleichungssystems und bei der Berechnung einer inversen Matrix ist die Anzahl der erforderlichen Multiplikationen

und Additionen geeignet. Wenn ein lineares Gleichungssystem $Au = r$ nach dem GAUSSschen Eliminationsverfahren gelöst wird, ist die Anzahl der erforderlichen Multiplikationen und Additionen je etwa $\frac{1}{3}n^3$. Wenn das Gleichungssystem nach dem JORDANschen Eliminationsverfahren gelöst und dabei gleichzeitig die inverse Matrix berechnet wird, ist die Anzahl der erforderlichen Multiplikationen und Additionen je etwa n^3 (s. BODEWIG [2]). Der Rechenaufwand bei der Anwendung einer der angegebenen Abschätzungen wächst also schon durch die Berechnung von A^{-1} etwa auf das dreifache des Rechenaufwandes an, der zur Lösung des linearen Gleichungssystems erforderlich ist. Die Frage, ob sich in einem speziellen Fall dieser erhöhte Rechenaufwand lohnt, kann hier natürlich nicht entschieden werden. Wenn dagegen die Aufgabe gestellt ist, eine Abschätzung für eine inverse Matrix $(A+C)^{-1}$ durchzuführen, wobei von den Elementen c_{ik} der Matrix C nur bekannt ist, daß ihre Beträge durch γ_{ik} beschränkt sind, tritt keine Erhöhung des Rechenaufwands ein. Dieser Fall kann durch n-malige Anwendung der angegebenen Abschätzungsformeln behandelt werden. Dabei sind die Vektoren r die Einheitsvektoren $\begin{pmatrix}1\\0\\\dots\\0\end{pmatrix}$, $\begin{pmatrix}0\\1\\\dots\\0\end{pmatrix}$, $\dots$ $\begin{pmatrix}0\\0\\\dots\\1\end{pmatrix}$, die Vektoren u sind die Spaltenvektoren der Matrix A^{-1}, und es ist $\sigma_i = 0$ $(i=1,2,\dots,n)$. In diesem Fall muß A^{-1} sowieso berechnet werden.

2. Die Berechnung von Schranken nach einer der Formeln (8.8), (8.11), (8.12), (8.13) oder (8.8*), (8.11*), (8.12*), (8.13*). Wenn alle $\gamma_{ik} = \gamma$ sind und die Formeln (8.8*), (8.11*), (8.12*), (8.13*) angewendet werden, sind zunächst die Summen $\beta_i = \sum |b_{ij}|$ und $\sum |u_k|$ zu berechnen. Außerdem kommen, je nach der verwendeten Formel, einige Additionen, Multiplikationen, Berechnungen von Quadratwurzeln, schließlich eine Division hinzu. Der hierfür erforderliche Rechenaufwand ist für größere n klein gegenüber dem Rechenaufwand bei der Lösung des Gleichungssystems und bei der Berechnung der inversen Matrix. Wenn dagegen die Abschätzung nach (8.8), (8.11) oder (8.12) durchgeführt wird, müssen

im wesentlichen drei oder vier Multiplikationen einer n-reihigen quadratischen Matrix mit einem Vektor durchgeführt werden. Das erfordert etwa je n^2 Multiplikationen und Additionen. Bei der Anwendung von (8.13) muß unter anderem das Produkt zweier n-reihiger quadratischer Matrizen gebildet werden, wofür etwa n^3 Multiplikationen und Additionen erforderlich sind.

Es soll nun eine Abschätzung hergeleitet werden, bei der man nicht A^{-1} berechnen muß, die aber für größere n im allgemeinen zu sehr viel gröberen Abschätzungen führt als die bisher beschriebenen Methoden. Beim GAUSSschen Eliminationsverfahren in der Form des Schemas von BANACHIEWICZ (s. BODEWIG [2]) sei die Matrix A in ein Produkt von zwei Dreiecksmatrizen A_1 und A_2 zerlegt worden:

$$A = A_1 \cdot A_2 = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ p_{21} & 1 & 0 & \cdots & 0 \\ p_{31} & p_{32} & 1 & \cdots & 0 \\ - & - & - & - & - \\ p_{n1} & p_{n2} & p_{n3} & \cdots & 1 \end{pmatrix} \cdot \begin{pmatrix} q_{11} & q_{12} & q_{13} & \cdots & q_{1n} \\ 0 & q_{22} & q_{23} & \cdots & q_{2n} \\ 0 & 0 & q_{33} & \cdots & q_{3n} \\ - & - & - & - & - \\ 0 & 0 & 0 & \cdots & q_{nn} \end{pmatrix}.$$

Die Zahlen in der Hauptdiagonale von A_2, nämlich $q_{11}, q_{22}, \ldots, q_{nn}$, seien alle von Null verschieden.

$\varsigma = \begin{pmatrix} \varsigma_1 \\ \varsigma_2 \\ \cdots \\ \varsigma_n \end{pmatrix}$ und $\xi = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \cdots \\ \xi_n \end{pmatrix}$ seien zwei Vektoren, die der Gleichung $A\xi = \varsigma$ genügen. Man kann Schranken ζ_i für $|\xi_i|$ berechnen, wenn man zunächst $\eta_1, \eta_2, \ldots, \eta_n$ nach der Vorschrift

$$\eta_1 = |\varsigma_1| \,, \quad \eta_i = |\varsigma_i| + \sum_{k=1}^{i-1} |p_{ik}| \eta_k \qquad (i=2,3,\ldots,n)$$

und dann $\zeta_n, \zeta_{n-1}, \ldots, \zeta_2, \zeta_1$ nach der Vorschrift

$$\zeta_n = \frac{\eta_n}{|q_{nn}|} \,, \quad \zeta_i = \frac{\eta_i + \sum_{k=i+1}^{n} |q_{ik}| \zeta_k}{|q_{ii}|} \qquad (i=n-1,\ldots,2,1)$$

bildet, wenn man nämlich den Eliminationsprozeß zur Berechnung von ξ aus ς durchführt, dabei aber von allen auftretenden Zahlen die Beträge nimmt. Schranken für die Beträge $|b_{ij}|$ der Elemente von A^{-1} erhielte man, wenn man diesen Prozeß n-mal

durchführte und dabei als Vektor ϱ die Einheitsvektoren $\begin{pmatrix}1\\0\\\dots\\0\end{pmatrix}, \begin{pmatrix}0\\1\\\dots\\0\end{pmatrix}, \dots, \begin{pmatrix}0\\0\\\dots\\1\end{pmatrix}$ wählte. Wenn man nun mit der Summe dieser Einheitsvektoren, mit dem Vektor $\begin{pmatrix}1\\1\\\dots\\1\end{pmatrix}$, den beschriebenen Prozeß durchführt, erhält man Zahlen ξ_i, die Schranken für die Summen $\sum_{j=1}^{n} |b_{ij}|$ sind, also für die in den Abschätzungsformeln (8.8*) bis (8.13*) auftretenden Zahlen β_i:

$$\beta_i \leq \xi_i \,. \tag{8.14}$$

Der Rechenaufwand zur Bestimmung der ξ_i ist ebenso groß wie der zur Berechnung eines Lösungsvektors zu gegebener rechter Seite, wenn die Dreiecksmatrizen A_1, A_2 bekannt sind (je etwa n^2 Multiplikationen und Additionen). Man kann in den Formeln (8.8*) bis (8.13*) die Schranken ξ_i statt der β_i verwenden und erhält so Schranken für $\|v-u\|$ oder $|v_1-u_1|$, wenn die sich ebenfalls ergebenden Schranken für die LIPSCHITZ-Konstanten kleiner als 1 sind. Bei der Durchrechnung von Beispielen zeigt es sich, daß man so meistens wesentlich gröbere Abschätzungen erhält als bei der Rechnung mit den β_i. Es gibt aber auch Matrizen A, bei denen keine gröberen Schranken entstehen (z. B. Matrizen, für die alle Diagonalelemente q_{ii} von A_2 positiv und alle anderen q_{ik} und die p_{ik} negativ werden).

Es soll nun die schon in § 1 angedeutete Möglichkeit besprochen werden, als Operator Q^{-1} statt des zur FRECHETschen Ableitung T' inversen Operators $(T')^{-1}$ einen anderen Operator zu verwenden, statt A^{-1} also eine andere Matrix D, etwa die Matrix, die bei der Matrixinversion unter dem Einfluß der Abrundungsfehler aus A entsteht. Mit $Q^{-1} = D$ wird

$$\begin{aligned} Hz &= z - D\left[Az + C(u+z) - s\right] \\ &= (E - DA)z - D\left[C(u+z) - s\right] \,, \end{aligned} \tag{8.15}$$

wobei E die n-reihige Einheitsmatrix ist. Für $D \neq A^{-1}$ ist $E - DA$ nicht die Nullmatrix. Die Elemente von D sollen mit d_{ij}, die von $E - DA$ mit δ_{ij} bezeichnet werden. Für z_r, $z_R \in R$ wird nun

$$Hz_r - Hz_R = (E - DA - DC)(z_r - z_R) \,,$$

und für $z_0 = 0$ wird

$$z_1 = Hz_0 = -D(Cu - s) .$$

Man kann diese Gleichungen benutzen, um Abschätzungen herzuleiten, die (8.8), (8.11), (8.12) und (8.13) entsprechen. Insbesondere erhält man die (8.11) entsprechende Aussage: Wenn mit

$$\alpha_i = \frac{1}{\sum_j |d_{ij}| (\sum_k \gamma_{jk} |u_k| + \sigma_j)} \qquad (i=1,2,\ldots,n)$$

$$K = \max_i \left(\sum_k (|\delta_{ik}| + \sum_j |d_{ij}| \gamma_{jk}) \frac{\alpha_i}{\alpha_k} \right) < 1$$

ist, besitzt (8.6) eine Lösung v mit

$$|v_l - u_l| \leq \frac{1}{1-K} \left\{ \sum_j |d_{lj}| (\sum_k \gamma_{jk} |u_k| + \sigma_j) \right\} . \qquad (l=1,2,\ldots,n)$$

Die bei dieser Abschätzung verwendete Matrix D muß so gewählt werden, daß die Beträge $|\delta_{ik}|$ der Elemente von E - DA klein sind, jedenfalls so klein, daß $K < 1$ wird. D darf sich also von A^{-1} nicht zu sehr unterscheiden.

Die Abschätzung (8.11*) ohne den Faktor $\frac{1}{1-K_2} = \frac{1}{1 - \gamma \sum \beta_k}$, also die auf die in § 4 beschriebene Art hergeleitete unexakte Abschätzung, findet sich bei BODEWIG [2], S. 141 - 142.

Eine andere Abschätzung ist bei WITTMEYER [11] angegeben. Dort wird $\sqrt{\sum (v_i - u_i)^2}$ unter Benutzung der charakteristischen Zahl von kleinstem Betrag der Matrix A'A und der charakteristischen Zahl von größtem Betrag der Matrix C'C abgeschätzt. Statt dieser charakteristischen Zahlen kann man untere bzw. obere Schranken für ihre Beträge verwenden, die man aus den a_{ik} und γ_{ik} berechnen kann. Diese Abschätzung erfordert zwar weniger Rechenaufwand als die in dieser Arbeit hergeleitete, liefert aber im allgemeinen viel gröbere Schranken, vor allem, wenn man die einzelnen Differenzen $|v_l - u_l|$ abschätzen möchte.

§9. Abschätzung bei Matrix-Eigenwertaufgaben

Zunächst werden einige in diesem Abschnitt verwendete Bezeichnungen erklärt:

1. Wenn M eine n-reihige quadratische Matrix ist, soll der erste Spaltenvektor von M mit M^* und die durch Weglassen der ersten Spalte aus M entstehende Matrix von n Zeilen und n-1 Spalten mit M^{**} bezeichnet werden.
2. Die durch Zusammenfügen eines Spaltenvektors w von n Komponenten und einer Matrix W von n Zeilen und n-1 Spalten entstehende n-reihige quadratische Matrix soll mit (w,W) bezeichnet werden.
3. E sei die n-reihige Einheitsmatrix.

A und C seien nun zwei n-reihige quadratische Matrizen. Von den Elementen c_{ik} von C sei nur bekannt, daß ihre Beträge durch nichtnegative reelle Zahlen γ_{ik} beschränkt sind: $|c_{ik}| \leq \gamma_{ik}$ $(i,k=1,2,\ldots,n)$. Eine einfache charakteristische Zahl λ von A, also eine einfache Nullstelle des charakteristischen Polynoms $P(\lambda) = \det(A-\lambda E)$, und der zugehörige Eigenvektor seien bekannt. Dieser sei so normiert, daß seine erste Komponente 1 ist (das ist, nötigenfalls durch Umnumerierung, immer zu erreichen), es sei also

$$(A-\lambda E)\begin{pmatrix}1\\ u_2\\ \cdots\\ u_n\end{pmatrix} = 0 \ .$$

Es sollen hinreichende Bedingungen dafür angegeben werden, daß die Matrix A + C eine charakteristische Zahl μ mit einem zugehörigen Eigenvektor $\begin{pmatrix}1\\ v_2\\ \cdots\\ v_n\end{pmatrix}$ besitzt, bei der man Schranken für $|\mu-\lambda|$ und $|v_i-u_i|$ (für i=2,3,...,n) angeben kann. Man faßt zunächst λ und $u_2, u_3, \ldots, u_n$ zu einem Vektor $u = \begin{pmatrix}\lambda\\ u_2\\ \cdots\\ u_n\end{pmatrix}$, μ und $v_2, v_3, \ldots, v_n$ zu einem Vektor $v = \begin{pmatrix}\mu\\ v_2\\ \cdots\\ v_n\end{pmatrix}$ zusammen. Der BANACHsche Raum $R_1 = R_2 = R$ (vgl. § 3) sei der Raum der Vektoren

$x = \begin{pmatrix} \kappa \\ x_2 \\ \dots \\ x_n \end{pmatrix}$ mit der Normdefinition $\|x\| = \operatorname{Max}_{2 \leq i \leq n} (\alpha_1 |\kappa|, \alpha_i |x_i|)$,

also mit der Normdefinition, die sich bei der Abschätzung für lineare Gleichungssysteme als besonders günstig erwiesen hat. Die positiven Konstanten $\alpha_1, \alpha_2, \dots, \alpha_n$ sollen wieder so bestimmt werden, daß sich möglichst kleine Schranken ergeben.

T sei der durch $Tx = (A-\kappa E)\begin{pmatrix} 1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}$ definierte Operator. Tx soll umgeformt werden:

$$Tx = \left((A-\kappa E)^*, (A- E)^{**}\right)\begin{pmatrix} 1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} = A^* + \left(-\kappa E^*, (A-\kappa E)^{**}\right)\begin{pmatrix} 1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}$$

$$= A^* + \left(- E^*, (A-\kappa E)^{**}\right)\begin{pmatrix} \kappa \\ x_2 \\ \dots \\ x_n \end{pmatrix} = A^* + \left(- E^*, (A-\kappa E)^{**}\right)x \ .$$

S sei der durch

$$Sx = (A+C- E)\begin{pmatrix} 1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}$$

definierte Operator. Man kann Sx ebenso wie Tx umformen und erhält

$$Sx = (A+C)^* + \left(- E^*, (A+C- E)^{**}\right) x \ .$$

Als Operator Q soll wieder die FRECHETsche Ableitung von T verwendet werden.

Mit $z = \begin{pmatrix} \nu \\ z_2 \\ \dots \\ z_n \end{pmatrix}$ wird

$$T(x+z) = A^* + \left(- E^*, (A-(\kappa+\nu)E)^{**}\right)(x+z)$$

$$= Tx + \left(- E^*, (A-\kappa E)^{**}\right)z + (0^*, -\nu E^{**})x + (0^*, -\nu E^{**})z \ .$$

Daher wird

$$T'_{(x)}z = (-E^*, (A-\kappa E)^{**})z + (0^*, -\nu E^{**})x$$

$$= (-E^*, (A-\kappa E)^{**})z + \begin{pmatrix} 0 \\ -\nu x_2 \\ \dots \\ -\nu x_n \end{pmatrix}$$

$$= \left(-\begin{pmatrix} 1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}, (A-\kappa E)^{**}\right) z \ .$$

Wenn λ eine einfache charakteristische Zahl von A und $\begin{pmatrix} 1 \\ u_2 \\ \cdots \\ u_n \end{pmatrix}$ der zugehörige Eigenvektor ist, ist die Matrix

$$\left(-\begin{pmatrix} 1 \\ u_2 \\ \cdots \\ u_n \end{pmatrix}, (A - \lambda E)^{**} \right)$$

nichtsingulär. Zunächst erkennt man, daß $(A - \lambda E)^{**}$ den Rang $n-1$ hat. $A - \lambda E$ hat nämlich, da λ eine einfache charakteristische Zahl ist, den Rang $n-1$. Die erste Spalte $(A - \lambda E)^{*}$ von $(A - \lambda E)$ ist wegen $(A - \lambda E)\begin{pmatrix} 1 \\ u_2 \\ \cdots \\ u_n \end{pmatrix} = 0$ eine Linearkombination der Spalten von $(A - \lambda E)^{**}$. Daher haben $A - \lambda E$ und $(A - \lambda E)^{**}$ den gleichen Rang. Wäre nun $\left(-\begin{pmatrix} 1 \\ u_2 \\ \cdots \\ u_n \end{pmatrix}, (A - \lambda E)^{**} \right)$ singulär, so wäre $\begin{pmatrix} 1 \\ u_2 \\ \cdots \\ u_n \end{pmatrix}$ eine Linearkombination der Spaltenvektoren von $(A - \lambda E)^{**}$ und daher auch eine Linearkombination der Spaltenvektoren von $A - \lambda E$. Es gäbe einen Vektor t mit $(A - \lambda E)t = \begin{pmatrix} 1 \\ u_2 \\ \cdots \\ u_n \end{pmatrix}$, also außer dem Eigenvektor $\begin{pmatrix} 1 \\ u_2 \\ \cdots \\ u_n \end{pmatrix}$ einen Hauptvektor t zur charakteristischen Zahl λ, die dann nach dem folgenden Hilfssatz nicht mehr einfache charakteristische Zahl wäre.

Hilfssatz: Wenn eine Matrix A zu einer charakteristischen Zahl λ einen Eigenvektor u und einen Hauptvektor zweiter Stufe t besitzt, wenn also $Au = \lambda u$ und $At = \lambda t + u$ ist, ist λ keine einfache charakteristische Zahl von A.

Beweis: Der Beweis wird indirekt geführt. A kann durch Transformation mit einer nichtsingulären Matrix V in die JORDANsche Normalform überführt werden: $V^{-1}AV = J$. Wenn λ eine einfache charakteristische Zahl ist, kann man V so wählen, daß

$$J = \begin{pmatrix} \lambda & 0 \ldots 0 \\ 0 & \\ \cdots & J_1 \\ 0 & \end{pmatrix}.$$

ist, wobei J_1 eine (n-1)-reihige quadratische Matrix in JORDANscher Normalform ist, deren Hauptdiagonalelemente alle von λ verschieden sind. Der zur charakteristischen Zahl λ gehörende Eigenvektor der Matrix J ist $V^{-1}u$ wegen $V^{-1}AV(V^{-1}u) = \lambda(V^{-1}u)$. Andererseits erkennt man, daß dieser Eigenvektor (bis auf einen Faktor) der Vektor $\begin{pmatrix} 1 \\ 0 \\ \dots \\ 0 \end{pmatrix}$ sein muß (dies ist ein Eigenvektor von J; da λ eine einfache charakteristische Zahl ist, ist er bis auf einen Faktor eindeutig bestimmt). Ferner ist

$$V^{-1}AV(V^{-1}t) = J(V^{-1}t) = \lambda(V^{-1}t) + \begin{pmatrix} 1 \\ 0 \\ \dots \\ 0 \end{pmatrix} .$$

Das bedeutet, daß der aus $V^{-1}t$ durch Streichung der ersten Komponente entstehende Vektor v^* ein Eigenvektor der Matrix J_1 zur charakteristischen Zahl λ ist: $J_1v^* = \lambda v^*$. v^* ist nicht der Nullvektor, da sonst $V^{-1}t$ bis auf einen Faktor dem Vektor $V^{-1}u = \begin{pmatrix} 1 \\ 0 \\ \dots \\ 0 \end{pmatrix}$ gleich wäre, t also bis auf einen Faktor dem Vektor u gleich wäre. Andererseits kann die Matrix J_1 wegen ihrer oben angegebenen Eigenschaften nicht die charakteristische Zahl λ haben. Die Annahme, λ sei eine einfache charakteristische Zahl von A, führt also auf einen Widerspruch.

Da die Matrix $\left(-\begin{pmatrix} 1 \\ u_2 \\ \dots \\ u_n \end{pmatrix}, (A-\lambda E)^{**}\right)$ nichtsingulär ist, existiert der inverse Operator $(T'_{(u)})^{-1}$, und es ist

$$(T'_{(u)})^{-1}w = \left(-\begin{pmatrix} 1 \\ u_2 \\ \dots \\ u_n \end{pmatrix}, (A-\lambda E)^{**}\right)^{-1} w .$$

Die Matrix $\left(-\begin{pmatrix} 1 \\ u_2 \\ \dots \\ u_n \end{pmatrix}, (A-\lambda E)^{**}\right)^{-1}$ soll im folgenden mit B und ihre Elemente mit b_{ik} bezeichnet werden. Mit $Tu = 0$ erhält man

$$
\begin{aligned}
S(u+z) &= (A+C)^{*} + \left(-E^{*},(A+C-(\lambda+\nu)E)^{**}\right)(u+z)\\
&= C^{*} + \left(-E^{*},(A+C-(\lambda+\nu)E)^{**}\right)z + (O^{*},(C-\nu E)^{**})u\\
&= C^{*} + (O^{*},(C-\nu E)^{**})z + (O^{*},C^{**})u +\\
&\quad + \left(-\begin{pmatrix}1\\ u_2\\ \cdots\\ u_n\end{pmatrix},(A-\lambda E)^{**}\right)z,
\end{aligned}
$$

wenn man dieselbe Umformung wie bei $T'_{(u)}$ vornimmt. Nun wird

$$
\begin{aligned}
Hz &= z - (T'_{(u)})^{-1}S(u+z)\\
&= z - B\left\{C^{*}+(O^{*},(C-\nu E)^{**})z + (O^{*},C^{**})u + \left(-\begin{pmatrix}1\\ u_2\\ \cdots\\ u_n\end{pmatrix},(A-\lambda E)^{**}\right)z\right\}\\
&= -B\left\{C^{*}+(O^{*},(C-\nu E)^{**})z + (O^{*},C^{**})u\right\}.
\end{aligned}
$$

Mit $z_0 = 0$ wird

$$
z_1 = Hz_0 = -B(C^{*}+(O^{*},C^{**})u) = -BC\begin{pmatrix}1\\ u_2\\ \cdots\\ u_n\end{pmatrix},
$$

die Komponenten z_{1i} von z_1 sind also

$$
z_{1i} = -\sum_{j=1}^{n} b_{ij}\left(c_{j1}+\sum_{k=2}^{n} c_{jk}u_k\right) \qquad (i=1,2,\ldots,n).
$$

und man erhält mit

$$
\|z_1\| \leq \max_i \left(\alpha_i \sum_{j=1}^{n} |b_{ij}|\left(\gamma_{j1}+\sum_{k=2}^{n}\gamma_{jk}|u_k|\right)\right) = \tau \qquad (9.1)
$$

die nach § 5 zur Abschätzung erforderliche Schranke τ. Die Zahlen α_i sollen wieder so gewählt werden, daß

$$
M_1 = \max_i \left(\frac{\alpha_i}{\alpha_1}\sum_{j=1}^{n} |b_{ij}|\left(\gamma_{j1}+\sum_{k=2}^{n}\gamma_{jk}|u_k|\right)\right)
$$

für $1=1,2,\ldots,n$ möglichst klein wird. Zu diesem Zweck setzt man

$$
\alpha_i = \frac{1}{\sum_{j=1}^{n} |b_{ij}|\left(\gamma_{j1}+\sum_{k=2}^{n}\gamma_{jk}|u_k|\right)} \qquad (i=1,2,\ldots,n) \;..\; (9.2)
$$

Damit wird

$$
M_1 = \sum_{j=1}^{n} |b_{1j}|\left(\gamma_{j1}+\sum_{k=2}^{n}\gamma_{jk}|u_k|\right) \qquad (1=1,2,\ldots,n)\;.
$$

Wenn ein in (9.2) als Nenner auftretender Ausdruck Null sein sollte, kann man, wie bei der Abschätzung für lineare Gleichungssysteme, alle $\alpha_i = 1$ setzen.

Nun ist noch zu untersuchen, ob es eine LIPSCHITZ-Konstante K gibt, die kleiner als 1 ist und mit der

$$\|Hz_I - Hz_{II}\| \leq K \|z_I - z_{II}\| \tag{9.3}$$

ist, wenn $\|z_I\|$ und $\|z_{II}\|$ nicht größer als $\frac{\tau}{1-K}$ sind. Wenn man (9.1) benutzt, ist K so zu bestimmen, daß (9.3) für

$$\left.\begin{aligned} \left.\begin{matrix} |z_{I_l}| \\ |z_{II_l}| \end{matrix}\right\} &\leq \frac{1}{1-K} \operatorname{Max}_i \left(\frac{\alpha_i}{\alpha_1} \sum_{j=1}^{n} |b_{ij}| \left(\gamma_{j1} + \sum_{k=2}^{n} \gamma_{jk} |u_k| \right) \right) \\ &= \frac{M_1}{1-K} \qquad (l=2,3,\ldots,n) \\ \left.\begin{matrix} |\nu_I| \\ |\nu_{II}| \end{matrix}\right\} &\leq \frac{M_1}{1-K} \end{aligned}\right\} \tag{9.4}$$

erfüllt ist. Es ist

$$\begin{aligned} Hz_I - Hz_{II} &= -B\left\{(0^*, (C-\nu_I E)^{**}) z_I - (0^*, (C-\nu_{II} E)^{**}) z_{II}\right\} \\ &= -B\left\{(0^*, C^{**})(z_I - z_{II}) + (0^*, -E^{**})(\nu_I z_I - \nu_{II} z_{II})\right. \\ &= \left. -B\left\{(0^*, C^{**})(z_I - z_{II}) + (0^*, -E^{**})\left[(\nu_I - \nu_{II}) z_I + \nu_{II}(z_I - z_{II})\right]\right\}\right. \end{aligned}$$

und daher

$$\begin{aligned} \|Hz_I - Hz_{II}\| &\leq \operatorname{Max}_i \left\{ \alpha_i \left[|b_{i1}| \sum_{k=2}^{n} \gamma_{1k} |z_{I_k} - z_{II_k}| + \right.\right. \\ &\qquad \left.\left. + \sum_{j=2}^{n} |b_{ij}| \left(\sum_{k=2}^{n} \gamma_{jk} |z_{I_k} - z_{II_k}| + |\nu_I - \nu_{II}| |z_{I_j}| + \right.\right.\right. \\ &\qquad \left.\left.\left. + |\nu_{II}| |z_{I_j} - z_{II_j}| \right) \right] \right\} \\ &\leq K \|z_I - z_{II}\| \end{aligned}$$

mit

$$K = \operatorname{Max}_i \left\{ \alpha_i \left[|b_{i1}| \sum_{k=2}^{n} \frac{\gamma_{1k}}{\alpha_k} + \sum_{j=2}^{n} |b_{ij}| \left(\sum_{k=2}^{n} \frac{\gamma_{jk}}{\alpha_k} + \frac{|z_{I_j}|}{\alpha_1} + \frac{|\nu_{II}|}{\alpha_j} \right) \right] \right\} . \tag{9.5}$$

Wie bei der Abschätzung für algebraische Gleichungen soll auch hier eine Iterationsvorschrift angegeben werden, nach der man positive Zahlen ξ_l (l=1,2,...,n) berechnen kann, die die folgende Eigenschaft haben: Mit

$$K(\xi_1, \ldots, \xi_n) = \operatorname{Max}_i \left\{ \alpha_i \left[|b_{i1}| \sum_{k=2}^{n} \frac{\gamma_{1k}}{\alpha_k} + \sum_{j=2}^{n} |b_{ij}| \left(\sum_{k=2}^{n} \frac{\gamma_{jk}}{\alpha_k} + \frac{\xi_j}{\alpha_1} + \frac{\xi_1}{\alpha_j} \right) \right] \right\} \tag{9.6}$$

ist

$$\xi_l = \frac{M_1}{1 - K(\xi_1, \ldots, \xi_n)} \qquad (l=1,2,\ldots,n). \tag{9.7}$$

Man berechnet dazu, beginnend mit $\xi_1^{(0)} = 0$ oder $\xi_1^{(0)} = M_1$, die Folgen $\xi_1^{(\varrho)}$ nach der Vorschrift

$$\xi_1^{(\varrho+1)} = \frac{M_1}{1 - K(\xi_1^{(\varrho)}, \ldots, \xi_n^{(\varrho)})} \qquad (l=1,2,\ldots,n;\ \varrho=0,1,2,\ldots)$$

Wenn es Zahlen $\xi_1, \xi_2, \ldots, \xi_n$ gibt, für die (9.7) mit $K(\xi_1,\ldots,\xi_n) < 1$ erfüllt ist, konvergieren die Folgen der $\xi_1^{(\varrho)}$ gegen ξ_1, weil sie dann monoton und beschränkt sind. Das folgt wieder daraus, daß $K(\xi_1,\ldots,\xi_n)$ eine monoton nicht fallende Funktion von $\xi_1, \xi_2, \ldots, \xi_n$ und daß $\xi_1^{(1)} \geqq \xi_1^{(0)}$ ist. Wenn andererseits $\xi_1, \xi_2, \ldots, \xi_n$ die Grenzwerte der Folgen $\xi_1^{(\varrho)}$ sind und $K = K(\xi_1,\ldots,\xi_n) < 1$ ist, ist mit diesem K auch wegen (9.5) und (9.6) die Ungleichung (9.3) erfüllt, wenn $|z_{r_j}| \leqq \xi_j$ $(j=2,3,\ldots,n)$ und $|v_r| \leqq \xi_1$ ist. (9.3) ist also sicher dann erfüllt, wenn die z_{r_1}, z_{r_1}, v_r, v_r den Ungleichungen (9.4) genügen, die man jetzt wegen (9.7) auch

$$\left.\begin{matrix} |z_{r_1}| \\ |z_{r_1}| \end{matrix}\right\} \leqq \xi_1 \quad (l=2,3,\ldots,n) \qquad \left.\begin{matrix} |v_r| \\ |v_r| \end{matrix}\right\} \leqq \xi_1$$

schreiben kann. Nach § 5 besitzt dann die Operatorgleichung

$$Sv = (A + C - \mu E)\begin{pmatrix} 1 \\ v_2 \\ \cdots \\ v_n \end{pmatrix} = 0$$

eine Lösung, nämlich eine charakteristische Zahl μ und einen zugehörigen Eigenvektor mit

$$\left.\begin{aligned} &|\mu - \lambda| \leqq \xi_1 \\ &|v_l - u_l| \leqq \xi_1 \quad (l=2,3,\ldots,n). \end{aligned}\right\} \qquad (9.8)$$

Wenn alle $\gamma_{ik} = \gamma$ sind, wird aus (9.2)

$$\alpha_i = \frac{1}{\gamma\beta_i(1 + \sum_{k=2}^{n} |u_k|)} \qquad (9.2^*)$$

mit $\beta_i = \sum_{j=1}^{n} |b_{ij}|$, und aus (9.6) wird

$$K(\xi_1,\ldots,\xi_n) = \max_i \left\{ \gamma \sum_{k=1}^{n} \beta_k + \frac{1}{\beta_i} \sum_{j=2}^{n} |b_{ij}| (\xi_j\beta_1 + \xi_1\beta_j) \right\}. \qquad (9.6^*)$$

Das gefundene Ergebnis soll in einem Satz zusammengefaßt werden:

Satz: λ sei eine einfache charakteristische Zahl der Matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ - & - & - & - \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} .$$

Die Komponenten des zu λ gehörenden Eigenvektors seien $1, u_2, u_3, \ldots, u_n$. Dann ist die Matrix

$$\begin{pmatrix} -1 & a_{12} & a_{13} & \cdots & a_{1n} \\ -u_2 & a_{22}-\lambda & a_{23} & \cdots & a_{2n} \\ -u_3 & a_{32} & a_{33}-\lambda & \cdots & a_{3n} \\ - & - & - & - & - \\ -u_n & a_{n2} & a_{n3} & \cdots & a_{nn}-\lambda \end{pmatrix}$$

nichtsingulär. b_{ij} seien die Elemente der Inversen dieser Matrix. C sei eine n-reihige quadratische Matrix, von deren Elementen c_{ik} nur bekannt ist, daß $|c_{ik}| \leq \gamma_{ik}$ mit nichtnegativen reellen Zahlen γ_{ik} ist. Es sei

$$M_l = \sum_{j=1}^{n} |b_{lj}| \left(\gamma_{j1} + \sum_{k=2}^{n} \gamma_{jk} |u_k|\right) \qquad (l=1,2,\ldots,n).$$

Wenn die Folgen $\xi_l^{(\varrho)}$ $(l=1,2,\ldots,n;\varrho=0,1,2,\ldots)$, die nach der Vorschrift

$$\xi_l^{(0)} = 0 \quad \text{oder} \quad \xi_l^{(0)} = M_l ,$$

$$\xi_l^{(\varrho+1)} = \frac{M_l}{1 - K(\xi_1^{(\varrho)}, \ldots, \xi_n^{(\varrho)})} \qquad (\varrho = 0,1,2,\ldots)$$

mit $K(\xi_1^{(\varrho)}, \ldots, \xi_n^{(\varrho)}) =$

$$= \max_i \left\{ \frac{1}{M_i} \left[|b_{i1}| \sum_{k=2}^{n} \gamma_{1k} M_k + \sum_{j=2}^{n} |b_{ij}| \left(\sum_{k=2}^{n} \gamma_{jk} M_k + \xi_j^{(\varrho)} M_1 + \xi_1^{(\varrho)} M_j \right) \right] \right\}$$

gebildet sind, gegen positive Zahlen $\xi_1, \xi_2, \ldots, \xi_n$ mit $K(\xi_1, \ldots, \xi_n) < 1$ konvergieren, besitzt die Matrix $A + C$ eine charakteristische Zahl μ mit $|\mu - \lambda| \leq \xi_1$ und einen zu μ gehörenden Eigenvektor mit den Komponenten $1, v_2, v_3, \ldots, v_n$, wobei

$$|v_2 - u_2| \leq \xi_2 , \quad |v_3 - u_3| \leq \xi_3 , \quad \ldots, \quad |v_n - u_n| \leq \xi_n$$

ist.

§ 10. Abschätzung bei linearen Integralgleichungen 2. Art

Gegeben sei eine lineare FREDHOLMsche Integralgleichung 2. Art

$$v(x) - f(x) - \lambda\int_a^b K(x,\xi)v(\xi)d\xi = 0 \,. \qquad (10.1)$$

$f(x)$ sei eine für reelle Zahlen x aus dem Intervall $a \leqq x \leqq b$ definierte und dort stetige reellwertige Funktion; $K(x,\xi)$ sei für $a \leqq x \leqq b$, $a \leqq \xi \leqq b$ reellwertig definiert und stetig. λ sei ein reeller Parameter. Gesucht ist eine für $a \leqq x \leqq b$ stetige Funktion $v(x)$, die der Gleichung (10.1) genügt. Der Kern $K(x,\xi)$ heißt entartet, wenn er sich als Summe von endlich vielen Produkten je einer Funktion von x und einer Funktion von ξ, nämlich als $\sum_{i=1}^{n}\varphi_i(x)\psi_i(\xi)$ darstellen läßt. Die Lösung einer Integralgleichung (10.1) mit entartetem Kern kann leicht angegeben werden (s. etwa KANTOROWITSCH-KRYLOW [6]). Eine bekannte Methode zur Bestimmung einer Näherungslösung einer Integralgleichung (10.1) mit nicht entartetem Kern verläuft so, daß $K(x,\xi)$ durch einen entarteten Kern angenähert und die Integralgleichung gelöst wird, die aus (10.1) entsteht, wenn man $K(x,\xi)$ durch diesen entarteten Kern ersetzt. Es handelt sich hier um den in § 1 beschriebenen Fall des Übergangs von einer gegebenen Aufgabe zu einer Ersatzaufgabe. Nach dem Verfahren von § 5 soll eine Abschätzung für den Unterschied der Lösungen beider Aufgaben angegeben werden. Dabei soll gleich der allgemeine Fall behandelt werden, daß $K(x,\xi)$ durch einen anderen Kern $k(x,\xi)$ ersetzt wird, zu dem ein sogenannter „lösender Kern" $\gamma(x,\xi,\lambda)$ bekannt ist, und daß $f(x)$ durch eine Funktion $g(x)$ ersetzt wird. Der Fall, daß $k(x,\xi)$ ein entarteter Kern ist, wird dann als Spezialfall erscheinen. Die Operatorgleichung $Sv = \Theta_2$ sei also die Gleichung (10.1). Die Operatorgleichung $Tu = \Theta_2$ laute

$$u(x) - g(x) - \lambda\int_a^b k(x,\xi)u(\xi)d\xi = 0 \,. \qquad (10.2)$$

$g(x)$ und $k(x,\xi)$ seien Funktionen mit den Eigenschaften, die oben für $f(x)$ und $K(x,\xi)$ gefordert wurden. Außerdem sei eine Funktion $\gamma(x,\xi,\lambda)$ bekannt, mit der

$$u(x) = g(x) + \lambda\int_a^b \gamma(x,\xi,\lambda)g(\xi)d\xi$$

(für beliebige stetige Funktionen g(x)) ist, falls u(x) Lösung von (10.2) ist. Eine solche Funktion $\gamma(x,\xi,\lambda)$ heißt „lösender Kern" oder „Resolvente" des Kerns $k(x,\xi)$. Die Differenz $K(x,\xi) - k(x,\xi)$ soll mit $R(x,\xi)$ bezeichnet werden.

Zur Anwendung des Abschätzungsverfahrens von § 5 wird der BANACHsche Raum $R_1 = R_2 = R$ der für $a \leqq x \leqq b$ stetigen Funktionen w(x) mit der Normdefinition

$$\|w\| = \operatorname*{Max}_{a\leqq x\leqq b} (A(x)|w(x)|)$$

eingeführt, wobei A(x) eine für $a \leqq x \leqq b$ positiv reellwertige stetige Funktion ist. Eine Lösung u(x) von (10.2) sei bekannt. Als Operator Q wird wieder die FRECHETsche Ableitung von T verwendet. Wegen

$$T(u+z) = u(x) + z(x) - g(x) - \lambda\int_a^b k(x,\xi)[u(\xi)+z(\xi)]\,d\xi$$

$$= Tu + z(x) - \lambda\int_a^b k(x,\xi)z(\xi)d\xi$$

wird

$$T'_{(u)}z = T'z = z(x) - \lambda\int_a^b k(x,\xi)d\xi$$

(die FRECHETsche Ableitung hängt nicht von u ab). Die Gleichung

$$T'z = z(x) - \lambda\int_a^b k(x,\xi)z(\xi)d\xi = w(x)$$

besitzt nach der FREDHOLMschen Theorie sicher dann zu gegebenem w(x) eine Lösung $z(x) = T'^{-1}w(x)$, wenn λ nicht Eigenwert der homogenen Integralgleichung $z(x) - \lambda\int_a^b k(x,\xi)z(\xi)d\xi = 0$ ist; dann ist

$$z(x) = T'^{-1}w(x) = w(x) + \lambda\int_a^b \gamma(x,\xi,\lambda)w(\xi)d\xi\ .$$

Ferner ist

$$S(u+z) = u(x) + z(x) - f(x) - \lambda\int_a^b K(x,\xi)[u(\xi)+z(\xi)]\,d\xi$$

$$= z(x) - [f(x) - g(x)] - \lambda\int_a^b K(x,\xi)z(\xi)d\xi -$$

$$- \lambda\int_a^b [K(x,\xi) - k(x,\xi)]\,u(\xi)d\xi$$

$$= T'z - [f(x) - g(x)] - \lambda\int_a^b R(x,\xi)[u(\xi)+z(\xi)]\,d\xi\ .$$

Für den Operator H erhält man daher

$$Hz = z - T'^{-1}S(u+z)$$
$$= T'^{-1}\left\{[f(x) - g(x)] + \lambda\int_a^b R(x,\xi)[u(\xi)+z(\xi)]\,d\xi\right\}.$$

Mit $z_0 = 0$ wird

$$z_1 = Hz_0 = T'^{-1}\left\{[f(x) - g(x)] + \lambda\int_a^b R(x,\xi)u(\xi)d\xi\right\}$$
$$= f(x) - g(x) + \lambda\int_a^b R(x,\xi)u(\xi)d\xi +$$
$$+ \lambda\int_a^b \gamma(x,\xi,\lambda)\left[f(\xi) - g(\xi) + \lambda\int_a^b R(\xi,\eta)u(\eta)d\eta\right]d\xi$$

und daher

$$\|z_1\| = \max_{a\le x\le b}\left\{A(x)|z_1(x)|\right\}$$
$$\le \max_{a\le x\le b}\left\{A(x)\left[|f(x)-g(x)| + |\lambda|\left|\int_a^b R(x,\xi)u(\xi)d\xi\right| +\right.\right.$$
$$\left.\left. + |\lambda|\int_a^b|\gamma(x,\xi,\lambda)|\left(|f(\xi)-g(\xi)| + |\lambda|\left|\int_a^b R(\xi,\eta)u(\eta)d\eta\right|\right)d\xi\right]\right\}.$$

Wenn $|f(x) - g(x)| \le r_1(x)$ und $\left|\int_a^b R(x,\xi)u(\xi)d\xi\right| \le r_2(x)$ ist, lautet diese Ungleichung

$$\|z_1\| \le \max_{a\le x\le b}\left\{A(x)\left[r_1(x) + |\lambda|r_2(x) +\right.\right.$$
$$\left.\left. + |\lambda|\int_a^b|\gamma(x,\xi,\lambda)|\left(r_1(\xi) + |\lambda|r_2(\xi)\right)d\xi\right]\right\}. \qquad (10.3)$$

Zur Bestimmung einer LIPSCHITZ-Konstante K bildet man

$$Hz_I - Hz_{II} = T'^{-1}\left\{\lambda\int_a^b R(x,\xi)[z_I(\xi)-z_{II}(\xi)]\,d\xi\right\}$$
$$= \lambda\int_a^b\left\{R(x,\xi)[z_I(\xi)-z_{II}(\xi)] +\right.$$
$$\left. + \lambda\int_a^b \gamma(x,\xi,\lambda)R(\xi,\eta)[z_I(\eta)-z_{II}(\eta)]\,d\eta\right\}d\xi\,.$$

Dann ist

$$\|Hz_I - Hz_{II}\| \le \max_{a\le x\le b}\left\{A(x)|\lambda|\int_a^b\left[\frac{|R(x,\xi)|}{A(\xi)} +\right.\right.$$
$$\left.\left. + |\lambda|\int_a^b|\gamma(x,\xi,\lambda)|\frac{|R(\xi,\eta)|}{A(\eta)}d\eta\right]d\xi\right\}\|z_I - z_{II}\|. \qquad (10.4)$$

Nun soll zunächst $A(x)$ so gewählt werden, daß sich eine günstige Abschätzung ergibt. Wenn es gelingt, mit Hilfe der Ungleichung (10.4) eine positive reelle Zahl K, die kleiner als 1 ist, zu bestimmen, mit der $\|Hz_I - Hz_{II}\| \leq K\|z_I - z_{II}\|$ für $z_I, z_{II} \in R$ ist (es ist hier nicht nötig, nach einem von K abhängenden Bereich zu suchen, in dem die LIPSCHITZ-Bedingung erfüllt ist, da in (10.4) der Faktor von $\|z_I - z_{II}\|$ nicht von z_I und z_{II} abhängt), dann ist die Existenz einer Lösung $v(x)$ von (10.1) mit

$$|v(x) - u(x)| \leq \frac{1}{A(x)} \|v - u\| \leq \frac{1}{A(x)} \frac{\tau}{1-K}$$

nachgewiesen, wobei τ eine aus (10.3) gewonnene Schranke für $\|z_1\|$ ist. Wenn man wieder bei der Wahl einer möglichst günstigen Funktion $A(x)$ von dem Faktor $\frac{1}{1-K}$ absieht, ist $A(x)$ so zu bestimmen, daß

$$M(x) = \max_{a \leq t \leq b} \left\{ \frac{A(t)}{A(x)} \left[r_1(t) + |\lambda| r_2(t) + \right. \right.$$
$$\left. \left. + |\lambda| \int_a^b |\gamma(t,\xi,\lambda)| \left(r_1(\xi) + |\lambda| r_2(\xi) \right) d\xi \right] \right\}$$

möglichst klein wird. Für $t = x$ ist $\frac{A(t)}{A(x)} = \frac{A(x)}{A(x)} = 1$, daher ist

$$M(x) \geq r_1(x) + |\lambda| r_2(x) + |\lambda| \int_a^b |\gamma(x,\xi,\lambda)| \left(r_1(\xi) + |\lambda| r_2(\xi) \right) d\xi .$$

$M(x)$ ist für alle x zwischen a und b gleich dieser unteren Grenze für $M(x)$, wenn

$$A(x) = \frac{1}{r_1(x) + |\lambda| r_2(x) + |\lambda| \int_a^b |\gamma(x,\xi,\lambda)| \left(r_1(\xi) + |\lambda| r_2(\xi) \right) d\xi}$$

gesetzt wird. Wenn die mit diesem $A(x)$ aus (10.4) berechnete LIPSCHITZ-Konstante K kleiner als 1 ist, ergibt sich die Abschätzung

$$|v(x) - u(x)| \leq \frac{r_1(x) + |\lambda| r_2(x) + |\lambda| \int_a^b |\gamma(x,\xi,\lambda)| \left(r_1(\xi) + |\lambda| r_2(\xi) \right) d\xi}{1 - K} . \qquad (10.5)$$

Dies ist eine Verallgemeinerung einer bei KANTOROWITSCH-KRYLOW [6], S. 137, angegebenen Abschätzung.

Wenn $k(x,\xi)$ ein entarteter Kern ist, wenn also

$$k(x,\xi) = \sum_{i=1}^{n} \varphi_i(x)\,\psi_i(\xi)$$

ist und $u(x)$ Lösung der Gleichung

$$u(x) - g(x) - \lambda \int_a^b \sum \varphi_i(x)\,\psi_i(\xi)u(\xi)d\xi = 0$$

sein soll, muß

$$u(x) = g(x) + \lambda \sum \alpha_i \varphi_i(x) \qquad (10.6)$$

mit

$$\alpha_i = \int_a^b \psi_i(\xi)u(\xi)d\xi$$

$$= \int_a^b \psi_i(\xi)g(\xi)d\xi + \lambda \sum_j \alpha_j \int_a^b \psi_i(\xi)\varphi_j(\xi)d\xi \qquad (i=1,2,\ldots,n)$$

sein. Das ist ein lineares Gleichungssystem für die α_i. Wenn man

$$\int_a^b \psi_i(\xi)\varphi_j(\xi)d\xi = a_{ij} \qquad (i,j=1,2,\ldots,n)$$

$$\int_a^b \psi_i(\xi)g(\xi)d\xi = g_i \qquad (i=1,2,\ldots,n)$$

setzt, lautet das lineare Gleichungssystem

$$\alpha_i - \lambda \sum_j a_{ij}\alpha_j = r_i \qquad (i=1,2,\ldots,n). \qquad (10.7)$$

A sei die quadratische n-reihige Matrix der a_{ij}, E die n-reihige Einheitsmatrix. Das Gleichungssystem (10.7) ist eindeutig nach den α_i lösbar, wenn die Matrix $E - \lambda A$ nichtsingulär ist, wenn also $\frac{1}{\lambda}$ nicht charakteristische Zahl der Matrix A ist oder wenn $\lambda = 0$ ist. Dann existiert die inverse Matrix $(E - \lambda A)^{-1}$, die mit

$$B = \begin{pmatrix} b_{11} & b_{12} \cdots b_{1n} \\ - - - - - - - - \\ b_{n1} & b_{n2} \cdots b_{nn} \end{pmatrix}$$

bezeichnet werden soll. Die Lösung des linearen Gleichungssystems (10.7) ist

$$\alpha_i = \sum_j b_{ij}g_j = \sum_{j=1}^{n} b_{ij} \int_a^b \psi_j(\xi)g(\xi)d\xi \qquad (i=1,2,\ldots,n).$$

Nach (10.6) ist daher der lösende Kern zum Kern

$k(x,\xi) = \sum_{i=1}^{n} \varphi_i(x)\psi_i(\xi)$ durch $\gamma(x,\xi,\lambda) = \sum_{i,j=1}^{n} \varphi_i(x) b_{ij} \psi_j(\xi)$

gegeben. Die Abschätzung (10.5) lautet dann

$$|v(x) - u(x)| \leq \frac{M(x)}{1-K}$$

mit

$$M(x) = r_1(x) + |\lambda| r_2(x) + |\lambda| \sum_{i,j=1}^{n} |\varphi_i(x)| |b_{ij}| \int_a^b |\psi_j(\xi)| \left[r_1(\xi) + |\lambda| r_2(\xi)\right] d\xi .$$

Diese Abschätzung gilt, wenn die als obere Schranke für

$$\operatorname*{Max}_{a \leq x \leq b} \left\{ \frac{|\lambda|}{M(x)} \int_a^b \left[|R(x,\xi)| M(\xi) + |\lambda| \sum_{i,j=1}^{n} |\varphi_i(x)| |b_{ij}| |\psi_j(\xi)| \int_a^b |R(\xi,\eta)| M(\eta) d\eta \right] d\xi \right\}$$

bestimmte Zahl K kleiner als 1 ist.

§ 11. Abschätzung bei gewöhnlichen Differentialgleichungen 1. Ordnung (Ersetzung einer nichtlinearen durch eine lineare Differentialgleichung)

Es sei eine Anfangswertaufgabe bei einer gewöhnlichen Differentialgleichung 1. Ordnung gegeben. Die Differentialgleichung und die Anfangsbedingung seien

$$v'(x) = f(x,v(x)) , \tag{11.1}$$

$$v(x_0) = v_0 . \tag{11.2}$$

Dabei sei $f(x,v)$ eine in einem Bereich B der reellen x,v-Ebene definierte reellwertige stetige Funktion. Die partielle Ableitung $\frac{\partial f(x,v)}{\partial v}$ soll in B existieren und stetig sein. Der Bereich B sei die Menge der Punkte (x,v) mit

$$x_0 \leq x \leq x_1 , \quad p_0(x) \leq v \leq p_1(x),$$

wobei $p_0(x)$ und $p_1(x)$ im Intervall $x_0 \leq x \leq x_1$ definierte stetige Funktionen mit $p_0(x) < p_1(x)$ sind. Gesucht ist eine für $x_0 \leq x \leq x_1$ definierte Funktion v(x), die dort stetig differenzierbar ist, der Differentialgleichung (11.1) und der Anfangs-

›edingung (11.2) genügt und dabei die Eigenschaft hat, daß die Punkte (x,v(x)) in B liegen. Aus (11.1) und (11.2) folgt durch Integration die Gleichung

$$v(x) - v_0 - \int_{x_0}^{x} f(\xi, v(\xi))\,d\xi = 0 \ . \tag{11.3}$$

Umgekehrt folgen hieraus wegen der für v(x) und f(x,v) geforderten Bedingungen die Gleichungen (11.1) und (11.2). Statt der Gleichung (11.3), die als Operatorgleichung $Sv = \Theta_2$ aufgefaßt wird, soll eine Operatorgleichung $Tu = \Theta_2$ gelöst werden, nämlich die Gleichung

$$u(x) - v_0 - \int_{x_0}^{x} [\varphi(\xi)u(\xi) + \psi(\xi)]\,d\xi = 0 \ , \tag{11.4}$$

die aus (11.3) entsteht, wenn man die im allgemeinen in v nichtlineare Funktion f(x,v) durch eine lineare Funktion ersetzt. $\varphi(x)$ und $\psi(x)$ seien zwei für $x_0 \leqq x \leqq x_1$ stetige Funktionen. Die zu (11.4) äquivalente Differentialgleichung und Anfangsbedingung sind

$$u'(x) = \varphi(x)u(x) + \psi(x) \ , \tag{11.5}$$

$$u(x_0) = v_0 \ . \tag{11.6}$$

(11.1) ist also durch die inhomogene lineare Differentialgleichung (11.5) ersetzt worden. Die Lösung der Anfangswertaufgabe (11.5), (11.6) und damit der Gleichung (11.4) kann nach der Methode der "Variation der Konstanten" bestimmt werden. Man erhält

$$u(x) = \exp\left(\int_{x_0}^{x} \varphi(\xi)d\xi\right)\left[v_0 + \int_{x_0}^{x} \psi(\eta)\exp\left(-\int_{x_0}^{\eta} \varphi(\xi)d\xi\right)d\eta\right] \ . \tag{11.7}$$

Es sollen nun hinreichende Bedingungen dafür angegeben werden, daß (11.3) eine Lösung v(x) besitzt, deren Unterschied von u(x) man abschätzen kann. Zunächst wird der BANACHsche Raum $R_1 = R_2 = R$ der für $x_0 \leqq x \leqq x_1$ stetigen reellwertigen Funktionen y(x) mit der Normdefinition

$$\|y\| = \mathrm{Max}(|y(x)|)$$

eingeführt. (Hier und im folgenden soll unter Max(F(x)) das Maximum der Funktion F(x) im Intervall $x_0 \leqq x \leqq x_1$ verstanden werden.) Als Operator Q soll die FRECHETsche Ableitung von T

verwendet werden. Da

$$T(u+z) = u(x) + z(x) - v_0 - \int_{x_0}^{x} \left[\varphi(\xi)\left(u(\xi)+z(\xi)\right) + \psi(\xi)\right] d\xi$$

$$= Tu + z(x) - \int_{x_0}^{x} \varphi(\xi) z(\xi) d\xi$$

ist, wird

$$T'_{(u)} z = T'z = z(x) - \int_{x_0}^{x} \varphi(\xi) z(\xi) d\xi$$

(die FRECHETsche Ableitung hängt nicht von u ab). Wenn w(x) eine in $x_0 \leqq x \leqq x_1$ stetig differenzierbare Funktion ist, kann die Gleichung

$$T'z = w(x) = w(x_0) + \int_{x_0}^{x} w'(\xi) d\xi$$

in der Form

$$z(x) - w(x_0) - \int_{x_0}^{x} \left[\varphi(\xi) z(\xi) + w'(\xi)\right] d\xi = 0$$

geschrieben werden. Sie kann ebenso wie (11.4) gelöst werden. Man erhält

$$z(x) = T'^{-1} w(x) = \exp\left(\int_{x_0}^{x} \varphi(\xi) d\xi\right)\left[w(x_0) + \int_{x_0}^{x} w'(\eta) \exp\left(-\int_{x_0}^{\eta} \varphi(\xi) d\xi\right) d\eta\right].$$

Ferner wird, wenn u(x) die durch (11.7) gegebene Lösung von (11.4) ist,

$$S(u+z) = u(x) + z(x) - v_0 - \int_{x_0}^{x} f(\xi, u(\xi)+z(\xi)) d\xi$$

$$= z(x) + \int_{x_0}^{x} \left[\varphi(\xi) u(\xi) + \psi(\xi) - f(\xi, u(\xi)+z(\xi))\right] d\xi$$

$$= T'z + \int_{x_0}^{x} \left[\varphi(\xi)\left(u(\xi)+z(\xi)\right) + \psi(\xi) - f(\xi, u(\xi)+z(\xi))\right] d\xi .$$

Der Operator H ist daher durch

$$Hz = z - T'^{-1} S(u+z)$$

$$= - T'^{-1} \left\{ \int_{x_0}^{x} \left[\varphi(\xi)\left(u(\xi)+z(\xi)\right) + \psi(\xi) - f(\xi, u(\xi)+z(\xi))\right] d\xi \right\}$$

$$= - \exp\left(\int_{x_0}^{x} \varphi(\xi) d\xi\right) \int_{x_0}^{x} \left[\varphi(\eta)\left(u(\eta)+z(\eta)\right) + \psi(\eta) - \right.$$

$$\left. - f(\eta, u(\eta)+z(\eta))\right] \exp\left(-\int_{x_0}^{\eta} \varphi(\xi) d\xi\right) d\eta =$$

$$= -\int_{x_0}^{x} \left[\varphi(\eta)(u(\eta)+z(\eta)) + \psi(\eta) - f(\eta,u(\eta)+z(\eta))\right] \exp\left(\int_{\eta}^{x} \varphi(\xi)d\xi\right) d\eta$$

gegeben. Mit $z_0 = 0$ wird

$$z_1 = Hz_0 = -\int_{x_0}^{x} \left[\varphi(\eta)u(\eta) + \psi(\eta) - f(\eta,u(\eta))\right] \exp\left(\int_{\eta}^{x} \varphi(\xi)d\xi\right) d\eta$$

und somit

$$\|z_1\| = \mathrm{Max}\left\{\int_{x_0}^{x} \left|\varphi(\eta)u(\eta) + \psi(\eta) - f(\eta,u(\eta))\right| \exp\left(\int_{\eta}^{x} \varphi(\xi)d\xi\right) d\eta\right\}. \quad (11.8)$$

Ferner wird für z_I , $z_{II} \in R$

$$Hz_I - Hz_{II} = -\int_{x_0}^{x} \left[\varphi(\eta)(z_I(\eta)-z_{II}(\eta)) - f(\eta,u(\eta)+z_I(\eta)) + f(\eta,u(\eta)+z_{II}(\eta))\right] \exp\left(\int_{\eta}^{x} \varphi(\xi)d\xi\right) d\eta .$$

Wenn $z_I(x)$ und $z_{II}(x)$ so gewählt sind, daß alle Punkte $(x,u(x)+z_I(x))$ und $(x,u(x)+z_{II}(x))$ für $x_0 \leqq x \leqq x_1$ in B liegen, kann man den Mittelwertsatz der Differentialrechnung anwenden und erhält

$$Hz_I - Hz_{II} = -\int_{x_0}^{x} \left[\varphi(\eta) - \frac{\partial f(\eta,u(\eta)+\tilde{z}(\eta))}{\partial v}\right] (z_I(\eta)-z_{II}(\eta)) \exp\left(\int_{\eta}^{x} \varphi(\xi)d\xi\right) d\eta. \quad (11.9)$$

Für jedes η ist dabei $\tilde{z}(\eta)$ eine Stelle zwischen $z_I(\eta)$ und $z_{II}(\eta)$. Die Gleichung (11.9) und die folgenden Gleichungen sind nur dann sinnvoll, wenn $\tilde{z}(\eta)$ als Funktion von η so beschaffen ist, daß $\frac{\partial f(\eta,u(\eta)+\tilde{z}(\eta))}{\partial v}$ eine (etwa im RIEMANNschen Sinne) integrierbare Funktion ist. Man kann sogar zeigen, daß dies eine im offenen Intervall $x_0 < x < x_1$ stetige und an den Intervallendpunkten x_0 und x_1 links- bzw. rechtsseitig stetige Funktion ist. An einer Stelle η, an der $z_I(\eta) - z_{II}(\eta) \neq 0$ ist, ist nämlich $\frac{\partial f(\eta,u(\eta)+\tilde{z}(\eta))}{\partial v}$ als Quotient $\frac{f(\eta,u(\eta)+z_I(\eta)) - f(\eta,u(\eta)+z_{II}(\eta))}{z_I(\eta) - z_{II}(\eta)}$ zweier stetiger Funktionen bei von Null verschiedenem Nenner stetig. An einer Stelle η, an der $z_I(\eta) - z_{II}(\eta) = 0$ ist, ist $\tilde{z}(\eta) = z_I(\eta) = z_{II}(\eta)$ zu setzen. Dort ist dann $\tilde{z}(\eta)$ selbst stetig, weil es in jedem Intervall um η durch die stetigen Funktionen $\mathrm{Min}(z_I, z_{II})$ und $\mathrm{Max}(z_I, z_{II})$ von unten und von oben beschränkt ist und an der Stelle η mit diesen Funktionen übereinstimmt.

Aus (11.9) erhält man

$$\|Hz_I - Hz_{II}\| \leq \operatorname{Max}\left\{\int_{x_0}^{x} \left|\varphi(\eta) - \frac{\partial f(\eta, u(\eta)+\tilde{z}(\eta))}{\partial v}\right| \exp\left(\int_{\eta}^{x}\varphi(\xi)d\xi\right)d\eta\right\} \|z_I - z_{II}\|. \tag{11.10}$$

Wenn es gelingt, eine positive Zahl K zu bestimmen, mit der

$$\operatorname{Max}\left\{\int_{x_0}^{x} \left|\varphi(\eta) - \frac{\partial f(\eta, u(\eta)+\tilde{z}(\eta))}{\partial v}\right| \exp\left(\int_{x_0}^{x}\varphi(\xi)d\xi\right)d\eta\right\} \leq K$$

für $|\tilde{z}(\eta)| \leq \frac{\tau}{1-K}$ ist, wobei τ eine aus (11.8) bestimmte Schranke für $\|z_1\|$ ist, dann ist nachgewiesen, daß für $\|z_I\|, \|z_{II}\| \leq \frac{\tau}{1-K}$ die Ungleichung $\|Hz_I - Hz_{II}\| \leq K\|z_I - z_{II}\|$ erfüllt ist. Wenn dann noch alle Punkte $(\eta, u(\eta)+\zeta)$ mit $x_0 \leq \eta \leq x_1$ und $|\zeta| \leq \frac{\tau}{1-K}$ zu B gehören und wenn K kleiner als 1 ist, weiß man nach § 5, daß (11.3) und damit die Anfangswertaufgabe (11.1), (11.2) eine Lösung v besitzt, für die $|v(x) - u(x)| \leq \|v - u\| \leq \frac{\tau}{1-K}$ im Intervall $x_0 \leq x \leq x_1$ ist.

Wenn ein fester Wert x^* gegeben ist und $|v(x^*) - u(x^*)|$ abgeschätzt werden soll, wird man den Endpunkt x_1 des Intervalls $[x_0, x_1]$, das x^* enthalten muß, so wählen, daß man für $|v(x^*) - u(x^*)|$ eine möglichst kleine Schranke erhält. Oft - wie bei dem Beispiel in § 12 - ist $x_1 = x^*$ der günstigste Wert für x_1. Hat man dann verschiedene Werte x^*, so ist jeweils bei der Abschätzung das Intervall zu benutzen, das x^* als rechten Endpunkt besitzt. Auf diese Weise kann, wenn statt x^* wieder x geschrieben wird, $|v(x) - u(x)|$ statt durch eine Konstante durch eine von x abhängende Funktion abgeschätzt werden.

§ 12. Beispiele

Beispiel zu § 6

Die Funktion $t(x) = 1 - \frac{x^2}{2} + \frac{x^4}{24}$, die nach dem dritten Glied abgebrochene Taylorreihe der Funktion $s(x) = \cos x$, hat die Nullstellen $u = \pm\sqrt{6 \pm \sqrt{12}} = \pm\, 3{,}076\,378$, $\pm\, 1{,}592\,450$. Wären die Nullstellen von $s(x)$, nämlich $\pm\frac{\pi}{2}$, $\pm\frac{3\pi}{2}$, ..., nicht bekannt, so könnte man doch versuchen, auf die in § 6 beschriebene Weise Schranken für einige Nullstellen von $s(x)$ zu berechnen.

Für $u = \sqrt{6 - \sqrt{12}} = 1{,}592\,450$ ist $t'(u) = -\,0{,}919\,402$ und $s(u) = -\,0{,}021\,652$ und daher $|z_1| = \left|\frac{s(u)}{t'(u)}\right| = 0{,}023\,550 = \tau$.

Will man die Abschätzung nach § 6 durchführen, so muß auf jeden Fall $\left|\frac{t'(u) - s'(u+\zeta)}{t'(u)}\right| \leqq K$ für $|\zeta| \leqq \tau$ sein. Das Maximum von $\left|\frac{t'(u) - s'(u+\zeta)}{t'(u)}\right|$ in diesem Bereich ist 0,087 664. Es wird an der Stelle angenommen, an der $s'(u+\zeta) = -1$ ist. Auch in dem Bereich $|\zeta| \leqq \frac{\tau}{1 - 0{,}087\,664} = 0{,}025\,813$ ist die Ungleichung $\left|\frac{t'(u) - s'(u+\zeta)}{t'(u)}\right| \leqq 0{,}087\,664$ erfüllt. Also ist $K = 0{,}087\,664$ als LIPSCHITZ-Konstante verwendbar, und da $K < 1$ ist, existiert eine Nullstelle v von $s(x) = \cos x$ mit

$$|v - 1{,}592\,450| \leqq 0{,}025\,813$$

(nämlich $\frac{\pi}{2} = 1{,}570\,796$). Bei der anderen Nullstelle von $t(x)$, bei $u = \sqrt{6 + \sqrt{12}} = 3{,}076\,378$, ist $\left|\frac{t'(u) - s'(u)}{t'(u)}\right| > 1$, eine Abschätzung nach § 6 ist also sicher nicht möglich.

Beispiel zu § 7

Das Polynom

$t(x) = 1,1247x^5 - 1,4422x^4 - 2,1088x^3 + 2,5495x^2 - 2,9168x - 5,1485$

wird in drei Faktoren zerlegt: $t(x) = 1,1247 \cdot t_1(x) \cdot t_2(x) \cdot t_3(x)$ mit

$$t_1(x) = x - 2,1436\,9542\,,$$
$$t_2(x) = x^2 - 1,3532\,2877\,x + 1,7441\,7139\,,$$
$$t_3(x) = x^2 + 2,2146\,2670\,x + 1,2243\,1108\,.$$

Mit dem Faktor $t_1(x)$ hat $t(x)$ die einfache reelle Nullstelle 2,1436 9542 , mit dem Faktor $t_2(x)$ das Paar konjugiert komplexer Nullstellen $0,6766\,1438 \pm i \cdot 1,1341\,8004$ und mit dem Faktor $t_3(x)$ das Paar reeller, nahe beieinander liegender Nullstellen $-1,0645\,1405$ und $-1,1501\,1265$.

Die Zahlen γ_k, durch die angegeben wird, um wieviel sich die Koeffizienten von $s(x)$ und von $t(x)$ unterscheiden können, seien $\gamma_k = 0,00005$ $(k=0,1,\ldots,5)$; $t(x)$ kann also aus $s(x)$ durch Abrundung auf vier Stellen hinter dem Komma entstanden sein. Für $u = 2,1436\,9542$, die Nullstelle von $t_1(x)$, erhält man (mit den Bezeichnungen von § 7)

$$\frac{\sum \gamma_k |u|^k}{|\sum k a_k u^{k-1}|} = \frac{0,0041\,991\,0}{40,868704} = 0,0001\,0275$$

$$\xi_0 = 0,0001\,0275 \qquad \varphi(\xi_0) = 0,0010\,3895$$
$$f(\xi_0) = \xi_1 = 0,0001\,0286 \qquad \varphi(\xi_1) = 0,0010\,3985$$
$$f(\xi_1) = \xi_2 = 0,00010286 = \xi^*, K = \varphi(\xi^*) = 0,0010\,3985\,.$$

$s(x)$ besitzt also eine Nullstelle v mit

$$|v - 2,1436\,9542| \leq 0,0001\,0286\,.$$

Für den quadratischen Faktor $t_2(x)$ wird

$$T'^{-1}_{\binom{a}{b}} = \begin{pmatrix} 0,0661\,6716 & -0,0432\,6235 \\ 0,0754\,5696 & 0,1247\,1102 \end{pmatrix},$$

$\bar{A}_r = 0{,}0014\,5533$, $\bar{B}_r = 0{,}0011\,9489$,

$T_{11}\bar{A}_r + T_{12}\bar{B}_r = \beta = 0{,}0001\,4799$, $T_{21}\bar{A}_r + T_{22}\bar{B}_r = \alpha = 0{,}0002\,5883$.

Die Berechnung der ξ_ν und η_ν ergibt folgende Zahlen:

ν	ξ_ν	η_ν	$K(\xi_\nu, \eta_\nu)$
0	0,0001 4799	0,0002 5883	0,0022 2522
1	0,0001 4832	0,0002 5941	0,0022 3016
2	0,0001 4832	0,0002 5941	

$\xi^* = 0{,}0001\,4832$, $\eta^* = 0{,}0002\,5941$

$K = K(\xi^*, \eta^*) = 0{,}0022\,3016$.

$s(x)$ besitzt einen quadratischen Faktor $x^2 + a'x + b'$ mit

$$-1{,}3533\,7709 \leqq a' \leqq -1{,}3530\,8045\,,$$
$$1{,}7439\,1198 \leqq b' \leqq 1{,}74443080\,.$$

Also existiert ein Paar konjugiert komplexer Nullstellen $-\frac{a'}{2} \pm i\sqrt{b' - \left(\frac{a'}{2}\right)^2}$ von $s(x)$ mit

$$0{,}6765\,4024 \leqq -\frac{a'}{2} \leqq 0{,}6766\,8855\,,$$

$$\sqrt{1{,}7439\,1198 - \left(\frac{a'}{2}\right)^2} \leqq \sqrt{b' - \left(\frac{a'}{2}\right)^2} \leqq \sqrt{1{,}7444\,3080 - \left(\frac{a'}{2}\right)^2}.$$

In der komplexen Zahlenebene ist der Bereich, in dem die Nullstellen liegen können, der Durchschnitt des Streifens zwischen den zur imaginären Achse parallelen Geraden $\mathrm{Re}\, z = 0{,}6765\,4024$ und $\mathrm{Re}\, z = 0{,}6766\,8355$ mit dem Kreisring zwischen den Kreisen $|z|^2 = 1{,}7439\,1198$ und $|z|^2 = 1{,}7444\,3080$.

Bei dem quadratischen Faktor $t_3(x)$ ist für $\gamma_k = 0{,}00005$ noch eine Abschätzung für die einzelnen reellen Nullstellen möglich, z.B. gibt es eine Nullstelle v von $s(x)$ mit

$$|v - 1{,}1501\,12651| \leqq 0{,}0003\,0384\,.$$

Wenn man dagegen $\gamma_k = 0{,}001$ $(k=0,1,\ldots,5)$ setzt, konvergiert die Folge der $\varphi(\xi_\nu)$ nicht mehr gegen ein $K<1$:

$\xi_0 = 0{,}0060\,5522$ $\qquad$ $\varphi(\xi_0) = 0{,}3405\,6572$

$\xi_1 = 0{,}0091\,8245$ $\qquad$ $\varphi(\xi_1) = 0{,}5114\,0214$

$$\xi_2 = 0{,}0123\,9305 \qquad \varphi(\xi_2) = 0{,}6888\,2506$$
$$\xi_3 = 0{,}0194\,5921 \qquad \varphi(\xi_3) = 1{,}1695\,7172 > 1 \ .$$

Man kann also nichts mehr über eine Nullstelle von s(x) in der Nähe von u = 1,1501 1265 aussagen; man kann aber von dem quadratischen Faktor $t_3(x) = x^2 + 2{,}2146\,2670\,x + 1{,}2243\,1108$ ausgehen und die Abschätzungsmethode von § 7 anwenden. Es wird

$$T'^{-1}_{\binom{a}{b}} = \begin{pmatrix} -\,0{,}0137\,1531 & 0{,}0677\,3777 \\ -\,0{,}0829\,3211 & 0{,}1362\,9857 \end{pmatrix},$$

$$\bar{A}_r = 0{,}0691\,9610 \ , \qquad \bar{B}_r = 0{,}0323\,7679 \ ,$$

$$T_{11}\bar{A}_r + T_{12}\bar{B}_r = \beta = 0{,}0031\,4218 \ ,$$

$$T_{21}\bar{A}_r + T_{22}\bar{B}_r = \alpha = 0{,}0101\,5149 \ ,$$

ν	ξ_ν	η_ν	$K(\xi_\nu,\eta_\nu)$
0	0,0031 4218	0,0101 5149	0,0967 7123
1	0,0034 7884	0,0112 3912	0,1059 8141
2	0,0035 1467	0,0113 5490	0,1069 6207
3	0,0035 1853	0,0113 6736	0,1070 6767
4	0,0035 1895	0,0113 6872	0,1070 7918
5	0,0035 1899	0,0113 6884	0,1070 8034
6	0,0035 1900	0,0113 6887	0,1070 8049
7	0,0035 1900	0,0113 6887	0,1070 8049

$$\xi^* = 0{,}0035\,1900 \ , \qquad \eta^* = 0{,}0113\,6887 \ ,$$
$$K = K(\xi^*,\eta^*) = 0{,}1070\,8049 \ .$$

s(x) besitzt einen quadratischen Faktor $x^2 + a'x + b'$ mit

$$|a' - 2{,}2146\,26701| \leqq 0{,}0035\,1900 \ ,$$
$$|b' - 1{,}2243\,11081| \leqq 0{,}0113\,6887 \ .$$

Für das Wurzelpaar $-\frac{a'}{2} \pm \sqrt{\left(\frac{a'}{2}\right)^2 - b'}$ von s(x) folgt hieraus:

$$-1{,}1090\,7285 \leqq -\frac{a'}{2} \leqq -1{,}1055\,5385 \ ,$$

$$\left|\sqrt{\left(\frac{a'}{2}\right)^2 - b'}\right| \begin{cases} \leqq 0{,}1307\,6842 & \text{für } \left(\frac{a'}{2}\right)^2 > b' \text{ (reelle Wurzel)} \\ \leqq 0{,}1158\,9064 & \text{für } \left(\frac{a'}{2}\right)^2 < b' \text{ (imag. Wurzel).} \end{cases}$$

In der komplexen Zahlenebene besteht der Bereich, in dem Wurzeln des quadratischen Faktors $x^2 + a'x + b'$ von s(x) liegen können, aus einem Intervall der reellen Achse und einem Be-

reich, der von den Parallelen zur imaginären Achse $\mathrm{Re}\, z = -1{,}1090\,7285$ und $\mathrm{Re}\, z = -1{,}1055\,5385$ und dem Kreis $|z|^2 = |b'_{max}| = 1{,}2356\,7995$ begrenzt wird.

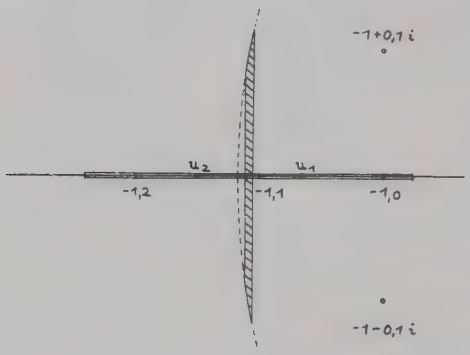

Abb. 2

In Abb. 2 ist dieser Bereich eingezeichnet, außerdem die beiden Nullstellen von $t_3(x)$.

Beispiel zu § 8

Das lineare Gleichungssystem

$$Au - r = \begin{pmatrix} 2{,}41 & -4{,}21 & 3{,}81 & -6{,}52 \\ -4{,}21 & 3{,}52 & 1{,}10 & 7{,}24 \\ 3{,}81 & 1{,}10 & -5{,}42 & -3{,}01 \\ -6{,}52 & 7{,}24 & -3{,}01 & 2{,}68 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} - \begin{pmatrix} 2{,}60 \\ 3{,}21 \\ -4{,}51 \\ 1{,}37 \end{pmatrix} = 0$$

hat die Lösung

$$u = \begin{pmatrix} 0{,}5694\,7283 \\ 1{,}4385\,4559 \\ 1{,}6192\,8831 \\ -0{,}1709\,1395 \end{pmatrix} .$$

Es sollen mit $\gamma = \varepsilon = 0{,}005$ Schranken für $\|v - u\|$ bzw. $|v_1 - u_1|$ nach den Formeln (8.8*), (8.11*), (8.12*) und (8.13*) berechnet werden. Es wird

$\beta_1 = 1{,}4503\,0872$, $\beta_3 = 1{,}1719\,3645$,

$\beta_2 = 1{,}7599\,4176$, $\beta_4 = 0{,}2917\,9868$,

$$\gamma \cdot \sum |u_k| + \sigma = 0{,}0239\,9110 .$$

Daraus ergibt sich nach (8.8*)

$$|v_1 - u_1| \leqq \|v - u\| \leqq 0{,}0437\,6336 \quad (l=1,2,3,4)$$

mit $K_1 = n\gamma \operatorname{Max}_i(\beta_i) = 0{,}0351\,9884$.

Nach (8.11*) erhält man

$$|v_1 - u_1| \leqq 0{,}0356\,2711 , \quad |v_3 - u_3| \leqq 0{,}0287\,8884 ,$$
$$|v_2 - u_2| \leqq 0{,}0432\,3330 , \quad |v_4 - u_4| \leqq 0{,}0071\,6809$$

mit $K_2 = \gamma \cdot \sum \beta_i = 0{,}0233\,6993$.

Nach (8.12*) wird

$$\sum |v_i - u_i| = \|v - u\| \leqq 0{,}1148\,1734$$

mit $K_3 = K_2 = 0{,}0233\,6993$, und nach (8.13*) erhält man

$$\sqrt{\sum (v_i - u_i)^2} = \|v - u\| \leqq 0{,}0635\,5083$$

mit $K_4 = \gamma \sqrt{n \sum \beta_i^2} = 0{,}0258\,0576$.

Für das hier behandelte lineare Gleichungssystem, das aus [14] übernommen ist, ist dort eine Abschätzung nach einer von COLLATZ und BARTSCH [4] angegebenen Verbesserung der Methode von WITTMEYER [11] durchgeführt, wobei sich

$$\sqrt{\sum (v_i - u_i)^2} \leqq 0{,}1585\,015$$

ergibt. Die kleinsten Schranken für $|v_1 - u_1|$ erhält man nach (8.11*). Wie gut diese Schranken sind, erkennt man, wenn man die linearen Gleichungssysteme

$$\begin{pmatrix} 2{,}405 & -4{,}215 & 3{,}805 & -6{,}515 \\ -4{,}215 & 3{,}515 & 1{,}095 & 7{,}245 \\ 3{,}805 & 1{,}095 & -5{,}425 & -3{,}005 \\ -6{,}515 & 7{,}245 & -3{,}005 & 2{,}675 \end{pmatrix} \begin{pmatrix} u_1^{(1)} \\ u_2^{(1)} \\ u_3^{(1)} \\ u_4^{(1)} \end{pmatrix} = \begin{pmatrix} 2{,}605 \\ 3{,}215 \\ -4{,}505 \\ 1{,}365 \end{pmatrix} ,$$

$$\begin{pmatrix} 2{,}405 & -4{,}215 & 3{,}805 & -6{,}515 \\ -4{,}215 & 3{,}515 & 1{,}095 & 7{,}245 \\ 3{,}805 & 1{,}095 & -5{,}425 & -3{,}005 \\ -6{,}525 & 7{,}235 & -3{,}015 & 2{,}685 \end{pmatrix} \begin{pmatrix} u_1^{(2)} \\ u_2^{(2)} \\ u_3^{(2)} \\ u_4^{(2)} \end{pmatrix} = \begin{pmatrix} 2{,}605 \\ 3{,}215 \\ -4{,}505 \\ 1{,}375 \end{pmatrix} ,$$

$$\begin{pmatrix} 2{,}415 & -4{,}205 & 3{,}815 & -6{,}525 \\ -4{,}215 & 3{,}515 & 1{,}095 & 7{,}245 \\ 3{,}815 & 1{,}105 & -5{,}415 & -3{,}015 \\ -6{,}515 & 7{,}245 & -3{,}005 & 2{,}675 \end{pmatrix} \begin{pmatrix} u_1^{(3)} \\ u_2^{(3)} \\ u_3^{(3)} \\ u_4^{(3)} \end{pmatrix} = \begin{pmatrix} 2{,}595 \\ 3{,}215 \\ -4{,}515 \\ 1{,}365 \end{pmatrix}$$

löst, deren Koeffizienten und rechte Seiten sich um nicht mehr als 0,005 von den entsprechenden a_{ik} und r_i unterscheiden, für die also (8.11*) noch gelten muß. Es wird

$$|u_1^{(1)} - u_1| = \underline{0{,}0355\,5574}\,, \qquad |u_3^{(1)} - u_3| = \underline{0{,}0287\,3109}\,,$$
$$|u_2^{(1)} - u_2| = 0{,}0405\,9311\,, \qquad |u_4^{(1)} - u_4| = 0{,}0000\,3954\,,$$
$$|u_1^{(2)} - u_1| = 0{,}0295\,3970\,, \qquad |u_3^{(2)} - u_3| = 0{,}0275\,5299\,,$$
$$|u_2^{(2)} - u_2| = \underline{0{,}0431\,4591}\,, \qquad |u_4^{(2)} - u_4| = 0{,}0046\,0003\,,$$
$$|u_1^{(3)} - u_1| = 0{,}0028\,0660\,, \qquad |u_3^{(3)} - u_3| = 0{,}0034\,3201\,,$$
$$|u_2^{(3)} - u_2| = 0{,}0098\,3099\,, \qquad |u_4^{(3)} - u_4| = \underline{0{,}0069\,6694}\,.$$

Die durch Unterstreichen hervorgehobenen Werte liegen sehr nahe an den aus (8.11*) berechneten Schranken. Wenn man (8.11*) ohne den Faktor $\frac{1}{1-K_2} = \frac{1}{1-\gamma\sum\beta_k}$, also die nicht exakte Abschätzung nach § 4 verwendet, erhält man

$$\beta_1(\gamma\sum|u_k|+\sigma) = 0{,}0347\,9450\,, \qquad \beta_3(\gamma\sum|u_k|+\sigma) = 0{,}0281\,1605\,,$$
$$\beta_2(\gamma\sum|u_k|+\sigma) = 0{,}0422\,2294\,, \qquad \beta_4(\gamma\sum|u_k|+\sigma) = 0{,}0070\,0057\,.$$

Außer der letzten werden alle diese Zahlen bei den obigen Beispielen überschritten.

Wenn man nach (8.14) Schranken ζ_i für β_i berechnet, findet man

$$\beta_1 \leqq 16{,}3040\,2773\,, \qquad \beta_3 \leqq 2{,}1282\,9406\,,$$
$$\beta_2 \leqq 5{,}9049\,1845\,, \qquad \beta_4 \leqq 0{,}8165\,9506\,.$$

Nach (8.11*) wird dann $K_2 \leqq 0{,}1257\,6918$ und

$$|v_1 - u_1| \leqq 0{,}4581\,3014\,, \qquad |v_3 - u_3| \leqq 0{,}0598\,0336\,,$$
$$|v_2 - u_2| \leqq 0{,}1659\,2349\,, \qquad |v_4 - u_4| \leqq 0{,}0229\,4567\,.$$

Die so berechneten Schranken für $|v_i - u_i|$ sind ersichtlich sehr viel größer als die oben angegebenen.

Wenn man A^{-1} nicht, wie das oben geschehen ist, auf 8 Stellen hinter dem Komma berechnet, sondern mit geringerer Genauigkeit, erhält man z. B.

$$A^{-1} \approx D = \begin{pmatrix} 0,355 & 0,544 & 0,428 & -0,123 \\ 0,544 & 0,673 & 0,490 & 0,052 \\ 0,428 & 0,490 & 0,229 & -0,024 \\ -0,123 & 0,052 & -0,024 & -0,093 \end{pmatrix}.$$

Man muß dann diese Ungenauigkeit bei der Abschätzung berücksichtigen und findet mit der am Ende von § 8 beschriebenen Methode $K = 0,0395\,5744$ (diese LIPSCHITZ-Konstante ist wesentlich größer als $K_2 = 0,0233\,6993$) und

$$|v_1 - u_1| \leqq 0,0362\,1987\,, \qquad |v_3 - u_3| \leqq 0,0292\,5066\,,$$
$$|v_2 - u_2| \leqq 0,0439\,3844\,, \qquad |v_4 - u_4| \leqq 0,0072\,9393\,.$$

Beispiel zu § 9

Als Matrix A wird dieselbe Matrix wie bei dem Beispiel zu § 8 gewählt:

$$A = \begin{pmatrix} 2,41 & -4,21 & 3,81 & -6,52 \\ -4,21 & 3,52 & 1,10 & 7,24 \\ 3,81 & 1,10 & -5,42 & -3,01 \\ -6,52 & 7,24 & -3,01 & 2,68 \end{pmatrix}.$$

Eine charakteristische Zahl von A ist $\lambda = 15,4623\,786$, der zu λ gehörende normierte Eigenvektor hat die Komponenten

$$u_1 = 1\,, \qquad u_3 = 0,2958\,695\,,$$
$$u_2 = -1,0306\,462\,, \qquad u_4 = -1,1635\,116\,.$$

Mit $\gamma = 0,005$ sollen nach (9.2*) und (9.6*) Schranken für $|u - \lambda|$ und $|v_1 - u_1|$ berechnet werden. Man erhält

$$M_1 = \frac{1}{\alpha_1} = 0,0173\,829\,, \qquad M_3 = \frac{1}{\alpha_3} = 0,0011\,412\,,$$
$$M_2 = \frac{1}{\alpha_2} = 0,0024\,666\,, \qquad M_4 = \frac{1}{\alpha_4} = 0,0021\,870\,.$$

Wenn man mit $\xi_1^{(0)} = M_1$ beginnt, erhält man folgende $\xi_1^{(\varrho)}$:

$\xi_1^{(1)} = 0{,}0175\,538 \quad \xi_1^{(2)} = 0{,}0175\,542 \quad \xi_1^{(3)} = 0{,}0175\,544$

$\xi_2^{(1)} = 0{,}0024\,909 \quad \xi_2^{(2)} = 0{,}0024\,910 \quad \xi_2^{(3)} = 0{,}0024\,910$

$\xi_3^{(1)} = 0{,}0011\,524 \quad \xi_3^{(2)} = 0{,}0011\,524 \quad \xi_3^{(3)} = 0{,}0011\,524$

$\xi_4^{(1)} = 0{,}0022\,085 \quad \xi_4^{(2)} = 0{,}0022\,085 \quad \xi_4^{(3)} = 0{,}0022\,085\,.$

Für $\varrho = 4$ wird $\xi_1^{(4)} = \xi_1^{(3)}$ $(l=1,2,3,4)$. Da die Folgen konvergieren, ist nachgewiesen, daß die Matrix A + C eine charakteristische Zahl μ und einen zugehörigen Eigenvektor $\begin{pmatrix} 1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix}$ mit

$$|\mu - \lambda| \leqq \xi_1 = 0{,}0175\,544\,, \qquad |v_3 - u_3| \leqq \xi_3 = 0{,}0011\,524\,,$$
$$|v_2 - u_2| \leqq \xi_2 = 0{,}0024\,910\,, \qquad |v_4 - u_4| \leqq \xi_4 = 0{,}0022\,085\,,$$

besitzt. Dabei ist $K = 0{,}0097\,561$.

Eine andere charakteristische Zahl von A ist $\lambda = 0{,}6977\,101$.

Der zugehörige normierte Eigenvektor hat die Komponenten

$$u_1 = 1\,, \qquad u_3 = 0{,}8776\,724\,,$$
$$u_2 = 1{,}2792\,799\,, \qquad u_4 = -\,0{,}0505\,440\,.$$

Die Folgen der $\xi_1^{(\varrho)}$ konvergieren gegen

$$\xi_1 = 0{,}0153\,335\,, \qquad \xi_3 = 0{,}0052\,702\,,$$
$$\xi_2 = 0{,}0052\,918\,, \qquad \xi_4 = 0{,}0041\,722\,.$$

Die Matrix A + C besitzt also eine charakteristische Zahl μ und einen zugehörigen Eigenvektor mit

$$|\mu - \lambda| \leqq \xi_1 \quad , \quad |v_1 - u_1| \leqq \xi_1 \quad (l=2,3,4).$$

Man kann die Abschätzung wiederholen, indem man eine andere Komponente als u_1 zu 1 normiert. Wenn u_2 zu 1 normiert wird, ergibt sich

$$|\mu - \lambda| \leqq 0{,}0153\,483\,, \qquad |v_1 - 0{,}7816\,8971| \leqq 0{,}0032\,366\,,$$
$$v_2 = 1\,, \qquad |v_3 - 0{,}6860\,6751| \leqq 0{,}0029\,585\,,$$
$$|v_4 + 0{,}0395\,0971| \leqq 0{,}0031\,538\,.$$

Wenn u_3 zu 1 normiert wird, findet man

$$|\mu - \lambda| \leqq 0{,}0153\,697\,, \qquad v_3 = 1\,,$$

$$|v_1 - 1{,}1393\,7731| \leqq 0{,}0068\,578\,,$$
$$|v_2 - 1{,}4575\,825\,| \leqq 0{,}0062\,943\,,$$
$$|v_4 + 0{,}5758\,6671| \leqq 0{,}0047\,381\,.$$

Wenn schließlich u_4 zu 1 normiert wird, erhält man

$$|\mu - \lambda| \leqq 0{,}0218\,696\,, \qquad v_4 = 1\,,$$

$$|v_1 + 19{,}7847\,461| \leqq 2{,}3293\,02\,,$$
$$|v_2 + 25{,}3102\,291| \leqq 2{,}8787\,63\,,$$
$$|v_3 + 17{,}3645\,251| \leqq 2{,}0328\,43\,.$$

Beispiel zu § 10

Bei der Integralgleichung

$$Sv = v(x) - 1 - \int_0^1 \frac{v(\xi)}{x+\xi+1}\,d\xi = 0$$

soll der Kern $K(x,\xi) = \frac{1}{x+\xi+1}$ durch einen entarteten Kern ersetzt werden. Wenn man zwei Konstanten A und B so zu bestimmen sucht, daß

$$M = \max_{0 \leqq x+\xi \leqq 2} \left| \frac{1}{x+\xi+1} - [A(x+\xi)+B] \right|$$

möglichst klein wird, findet man

$$A = -\,0{,}3333\,333\,, \quad B = 0{,}9106\,836\,, \quad M = 0{,}0893\,164\,.$$

Mit

$$\varphi_1(x) = 1\,, \qquad \psi_1(\xi) = 0{,}9106\,836 - 0{,}3333\,333\,\xi\,,$$
$$\varphi_2(x) = -\,0{,}3333\,333\,x\,, \qquad \psi_2(\xi) = 1$$

wird

$$\sum_{i=1}^{2} \varphi_i(x)\,\psi_i(\xi) = A(x+\xi) + B\,.$$

Die Integralgleichung mit entartetem Kern, durch die $Sv = 0$ ersetzt wird, sei also

$$Tu = u(x) - 1 - \int_0^1 \left[1 \cdot (0{,}9106\,836 - 0{,}3333\,333\,\xi) + (-\,0{,}3333\,333\,x) \cdot 1\right] u(\xi)\,d\xi = 0.$$

Dann wird

$$a_{11} = \int_0^1 \psi_1(\xi)\varphi_1(\xi)d\xi = 0{,}7440\,169, \quad a_{12} = \int_0^1 \psi_1(\xi)\varphi_2(\xi)d\xi = -0{,}1147\,436,$$

$$a_{21} = \int_0^1 \psi_2(\xi)\varphi_1(\xi)d\xi = 1\ , \qquad a_{22} = \int_0^1 \psi_2(\xi)\varphi_2(\xi)d\xi = -0{,}1666\,667,$$

$$g_1 = \int_0^1 \psi_1(\xi)g(\xi)d\xi = 0{,}7440\,169, \quad g_2 = \int_0^1 \psi_2(\xi)g(\xi)d\xi = 1\ .$$

Die Matrix

$$E - \lambda A = \begin{pmatrix} 0{,}2559\,831 & 0{,}1147\,436 \\ -1 & 1{,}1666\,667 \end{pmatrix}$$

ist nichtsingulär und hat die Inverse

$$(E - \lambda A)^{-1} = \begin{pmatrix} 2{,}8221\,900 & -0{,}2775\,671 \\ 2{,}4190\,200 & 0{,}6192\,282 \end{pmatrix}\ .$$

Damit wird

$$\alpha_1 = 1{,}8221\,900\ , \qquad \alpha_2 = 0{,}8063\,400$$

und

$$u(x) = 2{,}8221\,900 - 0{,}8063\,400\,x\ .$$

Mit den Bezeichnungen von § 10 wird

$$r_1(x) = 0\ ,$$

$$\left|\int_0^1 R(x,\xi)u(\xi)d\xi\right| \leq 0{,}2160\,582 = r_2(x)\ ,$$

$$M(x) = 0{,}6826\,851 + 0{,}1392\,553\,x\ ,$$

$$K = 0{,}3109\,989\ .$$

Da $K < 1$ ist, existiert eine Lösung v von $Sv = 0$ mit

$$|v(x) - u(x)| \leq \frac{M(x)}{1-K} = 0{,}9908\,331 + 0{,}2021\,118\,x\ .$$

Beispiel zu § 11

Bei der Anfangswertaufgabe

$$v'(x) = \frac{1-x}{1+v(x)} , \qquad v(0) = 1 \tag{12.1}$$

soll die nichtlineare Differentialgleichung durch eine lineare Differentialgleichung ersetzt werden. Unter der Annahme, daß in der Nähe der Anfangsstelle $x = 0$ die Funktion $v(x) - 1$ sich wenig von Null unterscheidet, ersetzt man

$$f(x,v) = \frac{1-x}{1+v} = \frac{1-x}{2+(v-1)} = \frac{1}{2}\,\frac{1-x}{1+\frac{v-1}{2}}$$

durch

$$\frac{1}{2}(1-x)\left(1-\frac{u-1}{2}\right) = \frac{1}{4}(1-x)(3-u)$$

und erhält so die lineare Anfangswertaufgabe

$$u'(x) = \frac{1}{4}(1-x)(3-u(x)) , \quad u(0) = 1 . \tag{12.2}$$

Sie hat die Lösung

$$u(x) = 3 - 2e^{(x^2-2x)/8} .$$

Zunächst soll die Abschätzungsmethode von § 11 für Intervalle $0 \leq x \leq x_1$ mit $x_1 \leq 1$ durchgeführt werden. Nach (11.8) wird

$$\|z_1\| = \max_{0\leq x\leq x_1}\left\{\int_0^x \left|\frac{1}{2}(1-\eta)e^{(\eta^2-2\eta)/8} - \frac{1-\eta}{4-2e^{(\eta^2-2\eta)/8}}\right| \exp\left(-\int_\eta^x \frac{1-\xi}{4}\,d\xi\right) d\eta\right\}$$

$$\leq \max_{0\leq x\leq x_1}\left\{\frac{1}{2}\int_0^x (1-\eta)\left|e^{(\eta^2-2\eta)/8} - \frac{1}{2-e^{(\eta^2-2\eta)/8}}\right| d\eta\right\}.$$

Da die Funktion $g(\eta) = \left|e^{(\eta^2-2\eta)/8} - \frac{1}{2-e^{(\eta^2-2\eta)/8}}\right|$ für $0 \leq \eta \leq 1$ monoton nicht fallend ist (soweit man das aus den berechneten Werten erkennen kann), kann man abschätzen:

$$\|z_1\| \leq \left|e^{(x_1^2-2x_1)/8} - \frac{1}{2-e^{(x_1^2-2x_1)/8}}\right| \frac{2x_1 - x_1^2}{4} = \tau .$$

Bei der Berechnung der LIPSCHITZ-Konstante K nach (11.10) erhält man zunächst

$$\max_{0\le x\le x_1}\left\{\int_0^x\left|-\frac{1-\eta}{4}+\frac{1-\eta}{(4-2e^{(\eta^2-2\eta)/8+\tilde z(\eta)})^2}\right|\exp\left(-\int_\eta^x\frac{1-\xi}{4}\,d\xi\right)d\eta\right\}$$

$$\le \max_{0\le x\le x_1}\left\{\int_0^x(1-\eta)\left|\frac14-\frac{1}{(4-2e^{(\eta^2-2\eta)/8+\zeta})^2}\right|d\eta\right\}.$$

Man muß versuchen, K so zu bestimmen, daß dieser letzte Ausdruck für alle ζ mit $|\zeta|\le\frac{\tau}{1-K}$ höchstens gleich K ist. Das kann man dadurch erreichen, daß man wie in §7 und §9 eine Folge von Zahlen ζ_ν nach der Vorschrift

$$\zeta_0=\tau,\quad \zeta_{\nu+1}=\frac{\tau}{1-K(\zeta_\nu)}\qquad(\nu=0,1,2,\dots)$$

mit

$$K(\zeta_\nu)=\left(x_1-\frac{x_1^2}{2}\right)\left|\frac14-\frac{1}{(4-2e^{(x_1^2-2x_1)/8+\zeta_\nu})^2}\right|$$

berechnet. Wenn diese ζ_ν gegen ein ζ^* mit $K(\zeta^*)<1$ konvergieren, ist die Existenz einer Lösung v(x) von (12.1) für $0\le x\le x_1$ mit

$$|v(x)-u(x)|\le\zeta^*=\frac{\tau}{1-K(\zeta^*)}=\frac{\tau}{1-K}$$

nachgewiesen. In der folgenden Tabelle sind die Ergebnisse dieser Rechnungen für einige Werte x_1 zwischen 0 und 1 zusammengestellt.

x_1	$e^{(x_1^2-2x_1)/8}$	$g(x_1)$	$\tau=\zeta_0$
0,2	0,955 998	0,001 855	0,000 166
0,4	0,923 116	0,005 489	0,000 878
0,6	0,900 325	0,009 035	0,001 897
0,8	0,886 920	0,011 488	0,002 757
1,0	0,882 497	0,012 355	0,003 089

x_1	$K(\zeta_0)$	ζ_1	$K(\zeta_1)$
0,2	0,003 720	0,000 167	0,003 720
0,4	0,011 072	0,000 888	0,011 073
0,6	0,018 322	0,001 933	0,018 325
0,8	0,023 383	0,002 823	0,023 389
1,0	0,025 181	0,003 169	0,025 188

Da $\zeta_2=\zeta_1$ für alle x_1 wird, kann ζ_1 als ζ^* und $K(\zeta_1)$ als K verwendet werden. Da stets $K<1$ ist, gibt es eine Lösung v(x)

von (12.1) mit $|v(x) - u(x)| \leq \xi_1$. Die Schranken ξ_1 sind der obigen Tabelle zu entnehmen. Man kann die Lösung $v(x)$ von (12.1) auch explizit berechnen:

$$v(x) = -1 + \sqrt{4 + 2x - x^2} \quad .$$

Damit erhält man zum Vergleich mit den angegebenen Schranken die folgenden Werte:

x_1	$u(x_1)$	$v(x_1)$	$\lvert v(x_1) - u(x_1)\rvert$
0,2	1,088 005	1,088 061	0,000 056
0,4	1,153 767	1,154 066	0,000 299
0,6	1,199 351	1,200 000	0,000 649
0,8	1,226 159	1,227 106	0,000 947
1,0	1,235 006	1,236 068	0,001 062

Die Lösung $u(x)$ von (12.2) ist symmetrisch zur Stelle $x = 1$, da $u(x) = u(2-x)$ ist. Wenn $v(x)$ die Lösung von (12.1) für $0 \leq x \leq 1$ ist, deren Existenz soeben nachgewiesen ist, erkennt man, daß $\bar{v}(x) = v(2-x)$ für $1 \leq x \leq 2$ Lösung der nichtlinearen Differentialgleichung in (12.1) ist:

$$\bar{v}'(x) = -v'(2-x) = -\frac{1-(2-x)}{1+v(2-x)} = \frac{1-x}{1+\bar{v}(x)} \quad ,$$

und daß man wegen $v(1) = \bar{v}(1)$ und $v'(1) = \bar{v}'(1) = 0$ aus v und $\bar{v}$ eine Lösung $v(x)$ von (12.1) für $0 \leq x \leq 2$ zusammensetzen kann, die ebenfalls symmetrisch zur Stelle $x = 1$ ist. Aus Symmetriegründen gelten dann die für einige Stellen x_1 zwischen 0 und 1 angegebenen Abschätzungen auch für die Stellen $2 - x_1$ zwischen 1 und 2. Bei $x = 2$ ist $v = u$. Man kann sogar über $x = 2$ hinausgehen und versuchen $|v(x) - u(x)|$ für $x \geq 2$ abzuschätzen. Hier kann man wieder zu dem symmetrisch gelegenen Bereich $x \leq 0$ übergehen und findet dann folgende Ergebnisse:

x_1		τ	ξ^*	K	|v-u|
-0,2	2,2	0,000 373	0,000 375	0,006 814	0,000 123
-0,4	2,4	0,004 471	0,004 650	0,038 477	0,001 447
-0,6	2,6	0,023 041	0,026 566	0,132 694	0,007 328
-0,8	2,8	0,086 385	0,191 750	0,549 493	0,027 121
-1,0	3,0	0,284 882	-	-	0,090 017

Für $x = -1$ bzw. $x = 3$ wird schon $K(\xi_0) > 1$. Daher ist keine Abschätzung mehr möglich.

§13. ALGOL-Programme

Die Abschätzungsverfahren der §§7 - 9 für algebraische Gleichungen, lineare Gleichungssysteme und Matrix-Eigenwertaufgaben sind von besonderer Bedeutung, wenn sehr umfangreiche derartige Aufgaben zu lösen sind, bei denen die Ausgangswerte (Koeffizienten der Gleichung, Elemente der Matrizen) nur mit beschränkter Genauigkeit bekannt sind. Zur Lösung solcher umfangreichen Aufgaben werden heute vielfach elektronische Rechenanlagen eingesetzt. Über die Genauigkeit der Ergebnisse kann man zunächst wenig oder gar nichts aussagen. Es ist daher wünschenswert, Programme zur Abschätzung nach den in dieser Arbeit angegebenen Methoden zu haben. Einige solche Programme sind im folgenden zusammengestellt. Sie sind in der algorithmischen Sprache ALGOL in der Fassung von 1960 [1] abgefaßt. Zu diesen Programmen ist zu bemerken:

1. Die „basic symbols", die in [1] durch Fettdruck gekennzeichnet sind, werden hier (gemäß Fußnote 1 auf Seite 110 von [1]) durch Unterstreichen kenntlich gemacht.

2. Bei dem auf das Symbol comment folgenden erklärenden Text werden einige bei der „publication language" erlaubte Schreibweisen verwendet: Griechische Buchstaben, tiefgestellte Indizes u.a.

3. Die Programme könnten an manchen Stellen so abgeändert werden, daß der Aufwand an Speicherplatz und Rechenzeit geringer wird. Das hätte aber zur Folge, daß die Programme weniger übersichtlich werden. Es wurde versucht, zwischen den einander entgegenwirkenden Tendenzen, einerseits möglichst

schnelle und wenig Speicherplatz benötigende, andererseits möglichst übersichtliche Programme zu schreiben, eine vernünftige Zwischenlösung zu finden.

```
procedure Linearfaktorabschätzung (a, gamma, n, u, epsilon,
          zeta, stop);
value n, u, epsilon;
integer n;        real u, epsilon, zeta;
array a, gamma;   label stop;
```

comment Wenn u eine Nullstelle des Polynoms $a_n x^n + \ldots + a_1 x + a_0$ ist und s(x) ein Polynom ist, dessen Koeffizienten sich von den a_k höchstens um γ_k unterscheiden, versucht das Programm iterativ eine Zahl ζ zu berechnen. Wenn diese Iteration konvergiert, besitzt s(x) eine Nullstelle, die sich von u höchstens um ζ unterscheidet. Wenn die Iteration nicht konvergiert, ist keine solche Aussage über eine Nullstelle von s(x) möglich. Dann endet das Programm mit dem label stop. Eine Zahl ε ist anzugeben, durch die entschieden wird, daß die Iteration zu beenden ist, wenn zwei aufeinander folgende Glieder der gegen ζ konvergierenden Folge sich um weniger als ε unterscheiden;

```
begin real s1, s2, s3, s4, z1, zeta0, K, u1; integer j;
      s1 := s2 := 0;
      for j := n step -1 until 0 do
          s1 := gamma[j] + abs(u) × s1;
      for j := n step -1 until 1 do
          s2 := j × a[j] + u × s2;
      if s2 = 0 then go to stop;
      z1 := zeta0 := s1/abs(s2);
      Iteration: u1 := abs(u) + zeta0; s3 := s4 := 0;
                 for j := n step -1 until 1 do
                     s3 := j× gamma[j] + u1 × s3;
                 for j := n step -1 until 2 do
                     s4 := j×(j-1)×abs(a[j]) + u1 × s4;
                 K := (s3 + zeta0 × s4)/abs(s2);
                 if K ≥ 1 then go to stop;
                 zeta := z1/(1 - K);
```

```
                if zeta - zeta0 ≧ epsilon then
                   begin zeta0 := zeta; go to Iteration end
end Linearfaktorabschätzung


procedure Quadratfaktorabschätzung (a, gamma, n, alfa, beta,
          epsilon, zeta, eta, stop);
value n, alfa, beta, epsilon;
integer n;        real alfa, beta, epsilon, zeta, eta;
array a, gamma;   label stop;
```

comment Wenn das Polynom $a_n x^n + \ldots + a_1 x + a_o$ den quadratischen Faktor $x^2 + \alpha x + \beta$ enthält und s(x) ein Polynom ist, dessen Koeffizienten sich von den a_k höchstens um γ_k unterscheiden, versucht das Programm, iterativ zwei Zahlen ζ und η zu berechnen. Wenn diese Iteration konvergiert, besitzt s(x) einen quadratischen Faktor $x^2 + \alpha' x + \beta'$, dessen Koeffizienten α' und β' sich von α und β höchstens um ζ bzw. η unterscheiden. Wenn die Iteration nicht konvergiert, ist keine Aussage über einen quadratisc n Faktor von s(x) möglich. Dann endet das Programm mit dem label stop. Eine Zahl ε ist anzugeben, durch die entschieden wird, daß die Iteration zu beenden ist, wenn der Unterschied zweier aufeinander folgender Glieder sowohl der gegen ζ als auch der gegen η konvergierenden Folge kleiner als ε ist;

```
begin real Aa, Ab, Ba, Bb, Det, T11, T12, T21, T22, zeta0,
           eta0, beta1, alfa1, alfa0, beta0, Aa1, Ab1, Ba1,
           Bb1, Zw1, Zw2, Aaa, Aab, Abb, Baa, Bab, Bbb,
           M11, M12, M21, M22, K1, K2, K;
      integer j;  array q[0:if n ≧ 6 then n else 6];
      procedure Divisionsalgorithmus (p1, p2, u, v, n, l);
      real u, v;   integer n, l;   array p1, p2;
      begin integer i;
            for i := n step -1 until l do
                p2[i] := p1[i];
            for i := n step -1 until l+2 do
                begin p2[i-1] := p2[i-1] - u × p2[i];
                      p2[i-2] := p2[i-2] - v × p2[i] end
      end Divisionsalgorithmus;
```

```
Quadratfaktorabschätzung:
     if n<6 then
        for j := n+1 step 1 until 6 do q[j] := 0;
     Divisionsalgorithmus (a, q, alfa, beta, n, 1);
     Divisionsalgorithmus (q, q, alfa, beta, n, 2);
     Aa := alfa× q[3] - q[2];     Ab := - q[3];
     Ba := beta× q[3];            Bb := - q[2];
     Det := Aa × Bb - Ab × Ba;
     if Det = 0 then go to stop;
     T11 := abs(Bb/Det);     T12 := abs(Ab/Det);
     T21 := abs(Ba/Det);     T22 := abs(Aa/Det);
     Divisionsalgorithmus (gamma, q, -abs(alfa), -abs(beta),
                           n, 0);
     zeta0 := beta1 := T11 × q[1] + T12 × q[0];
     eta0 := alfa1 := T21 × q[1] + T22 × q[0];
     Iteration:
          alfa0 := abs(alfa) + zeta0; beta0 := abs(beta) + eta0;
          Divisionsalgorithmus (gamma,q,-alfa0,-beta0,n,1);
          Divisionsalgorithmus (q,q,-alfa0,-beta0,n,2);
          Aa1 := alfa0 × q[3] + q[2];  Ab1 := q[3];
          Ba1 := beta0 × q[3];         Bb1 := q[2];
          for j := 1 step 1 until n do q[j] := abs(a[j]);
          Divisionsalgorithmus (q,q,-alfa0,-beta0,n,1);
          Divisionsalgorithmus (q,q,-alfa0,-beta0,n,2);
          Zw1 := q[3];    Zw2 := q[4];
          Divisionsalgorithmus (q,q,-alfa0,-beta0,n,3);
          Aaa := 2×(q[3] + alfa0 × q[4]);
          Aab := 2 × q[4];    Abb := 2× q[5];
          Baa := 2 × beta0 × q[4];
          Bab := 2 × beta0 × q[5] + Zw1;
          Bbb := 2 × beta0 × q[6] + 2 × Zw2;
          M11 := zeta0 × Aaa + eta0 × Aab + Aa1;
          M12 := zeta0 × Aab + eta0 × Abb + Ab1;
          M21 := zeta0 × Baa + eta0 × Bab + Ba1;
          M22 := zeta0 × Bab + eta0 × Bbb + Bb1;
          K1 := (T11 × M11 + T12 × M21)
                + (alfa1/beta1) × (T11 × M12 + T12 × M22);
          K2 := (beta1/alfa1) × (T21 × M11 + T22 × M21) ·
                + (T21 × M12 + T22 × M22);
```

```
        K := if K1 > K2 then K1 else K2;
        if K ≧ 1 then go to stop;
        zeta := beta1/(1 - K);  eta := alfa1/(1 - K);
        if zeta - zeta0 ≧ epsilon  eta - eta0 ≧ epsilon then
           begin zeta0 := zeta;  eta0 := eta;
                 go to Iteration end
end Quadratfaktorabschätzung
```

```
procedure LGSAbschätzung (a, gamma, r, sigma, n, u, z, stop);
value n;
integer n;   label stop;
array a, gamma, r, sigma, u, z;
```

comment Wenn zwei lineare Gleichungssysteme

$$a_{i1}u_1 + \ldots + a_{in}u_n = r_i \quad (i=1,2,\ldots,n),$$

$$(a_{i1}+c_{i1})v_1 + \ldots + (a_{in}+c_{in})v_n = r_i + s_i \quad (i=1,2,\ldots,n)$$

gegeben und γ_{ik} Schranken für $|c_{ik}|$ und σ_i Schranken für $|s_i|$ sind, werden durch das Programm die u_k und Schranken z_k für $|v_k - u_k|$ nach (8.11) berechnet. Wenn die Berechnung der Schranken nicht möglich ist, endet das Programm mit dem label stop. Es wird vorausgesetzt, daß ein Unterprogramm zur Matrixinversion vorhanden ist, das eine n-reihige quadratische Matrix durch ihre Inverse ersetzt;

```
begin integer i, j, k, l;  real Summe, K;
      array Zw, m [1:n];
      Matrixinversion(a,n);
      for i := 1 step 1 until n do
          begin Summe := 0;
                for k := 1 step 1 until n do
                    Summe := Summe + a[i,k] × r[k];
                    u[i] := Summe end;
      for j := 1 step 1 until n do
          begin Summe := sigma[j];
                for k := 1 step 1 until n do
                    Summe := Summe + gamma[j,k] × abs(u[k]);
                Zw[j] := Summe end;
```

```
        for l := 1 step 1 until n do
            begin Summe := 0;
                  for j := 1 step 1 until n do
                      Summe := Summe + abs(a[l,j]) × Zw[j];
                  m[l] := Summe end;
        for j := 1 step 1 until n do
            begin Summe := 0;
                  for k := 1 step 1 until n do
                      Summe := Summe + gamma[j,k] × m[k];
                  Zw[j] := Summe end;
        K := 0;
        for i := 1 step 1 until n do
            begin Summe := 0;
                  for j := 1 step 1 until n do
                      Summe := Summe + abs(a[i,j]) × Zw[j];
                  Summe := Summe/m[i];
                  if Summe > K then K := Summe end;
        if K ≥ 1 then go to stop;
        for i := 1 step 1 until n do
            z[i] := m[i]/(1 - K);
end LGSAbschätzung
```

```
procedure Eigenwertabschätzung (a, gamma, n, lambda, u, zeta,
          epsilon, stop);
value n, u, epsilon, lambda;
integer n;                  real lambda, epsilon;
array a, gamma, u, zeta;    label stop;
```

comment λ sei eine einfache charakteristische Zahl der Matrix a, und u sei der zugehörige Eigenvektor, dessen erste Komponente u_1 zu 1 normiert sei. c sei eine Matrix, deren Elemente c_{ik} dem Betrage nach durch γ_{ik} beschränkt sind. Das Programm versucht, iterativ Zahlen $\zeta_1, \zeta_2, \ldots, \zeta_n$ zu berechnen. Wenn diese Iteration konvergiert, besitzt die Matrix a + c eine charakteristische Zahl μ mit $|\mu - \lambda| \leq \zeta_1$ und einen (wie u normierten) zugehörigen Eigenvektor v mit $|v_2 - u_2| \leq \zeta_2, \ldots, |v_n - u_n| \leq \zeta_n$. Wenn die Iteration nicht konvergiert, liefert das Programm keine Aussage über eine charakteristische Zahl

von a + c und endet dann mit dem label stop. Eine Zahl ε ist anzugeben, durch die entschieden wird, daß die Iteration abzubrechen ist, wenn bei keiner der gegen die ζ_i konvergierenden Folgen sich zwei aufeinander folgende Glieder um ε oder mehr unterscheiden. Es wird vorausgesetzt, daß ein Unterprogramm zur Matrixinversion vorhanden ist, das eine n-reihige quadratische Matrix durch ihre Inverse ersetzt;

```
begin integer i, j, k, l;     real Summe,K;
      array Zw, Zw1, M, zeta1 [1:n];
      for i := 1 step 1 until n do
          a[i,1] := - u[i];
      for i := 2 step 1 until n do
          a[i,i] := a[i,i] - lambda;
      Matrixinversion (a,n);
      for j := 1 step 1 until n do
          begin Summe := 0;
                for k := 1 step 1 until n do
                    Summe := Summe + gamma[j,k] × abs(u[k]);
                Zw[j] := Summe end;
      for l := 1 step 1 until n do
          begin Summe := 0;
                for j := 1 step 1 until n do
                    Summe := Summe + abs(a[l,j]) × Zw[j];
                M[l] := Summe end;
      for j := 1 step 1 until n do
          begin Summe := 0;
                for k := 1 step 1 until n do
                    Summe := Summe + gamma[j,k] × M[k];
                Zw[j] := Summe end;
      for i := 1 step 1 until n do
          begin Summe := 0;
                for j := 1 step 1 until n do
                    Summe := Summe + abs(a[i,j]) × Zw[j];
                Zw1[i] := Summe end;
      for i := 1 step 1 until n do
          zeta1[i] := M[i];
```

```
      Iteration: K := 0
                 for i := 1 step 1 until n do
                     begin Summe := 0;
                           for j := 2 step 1 until n do
                               Summe := Summe + abs(a[i,j]×
                               (zeta1[j]×M[1]+zeta1[1]×M[j]));
                           K1 := (Zw1[i]+Summe)/M[i];
                           if K1 > K then K := K1 end;
                 if K ≥ 1 then go to stop;
                 for l := 1 step 1 until n do
                     zeta[l] := M[l]/(1 - K);
                 for l := 1 step 1 until n do
                     if abs(zeta[l] - zeta1[l]) ≧ epsilon then
                        begin for i := 1 step 1 until n do
                              zeta1[i] := zeta[i];
                              go to Iteration end
end Eigenwertabschätzung
```

Literaturverzeichnis

[1] BACKUS, J.W. und andere: Report on the algorithmic language ALGOL 60. Numerische Mathematik 2 (1960), 106-136.

[2] BODEWIG, E.: Matrix calculus. 2. Auflage, Amsterdam (1959).

[3] COLLATZ, L.: Numerische Behandlung von Differentialgleichungen. 2. Auflage, Berlin, Göttingen, Heidelberg (1955).

[4] COLLATZ, L.: Zur Fehlerabschätzung bei linearen Gleichungssystemen.
BARTSCH, H.: Abschätzungen für die kleinste charakteristische Zahl einer positiv-definiten hermiteschen Matrix.
Z. angew. Math. Mech. 34 (1954),S. 71 - 74.

[5] HOUSEHOLDER, A.S.: Principles of numerical analysis. New York (1953).

[6] KANTOROWITSCH, L.A., W.I. KRYLOW: Näherungsmethoden der höheren Analysis, Berlin (1956).

[7] LJUSTERNIK, L.A., W.I. SOBOLEW: Elemente der Funktionalanalysis. Berlin (1956).

[8] SCHRÖDER, J.: Das Iterationsverfahren bei allgemeinerem Abstandsbegriff. Math. Z. 66 (1956), S. 111 - 116.

[9] SCHRÖDER, J.: Nichtlineare Majoranten beim Verfahren der schrittweisen Näherung. Arch. Math. 7 (1956), S. 471 - 484.

[10] WILLERS, F.A.: Methoden der praktischen Analysis. Berlin (1950).

[11] WITTMEYER, H.: Einfluß der Änderung einer Matrix auf die Lösung des zugehörigen Gleichungssystems, sowie auf die charakteristischen Zahlen und die Eigenvektoren.
Z. angew. Math. Mech. 16 (1936), S. 287 - 300.

[12] ZURMÜHL, R.: Matrizen. 2. Auflage, Berlin, Göttingen, Heidelberg (1958).

[13] ZURMÜHL, R.: Praktische Mathematik für Ingenieure und Physiker. 2. Auflage, Berlin, Göttingen, Heidelberg (1957)

[14] COLLATZ, L.: Numerische und Graphische Methoden. In "Handbuch der Physik" Bd. II, Berlin, Göttingen, Heidelberg (1955).

Lebenslauf

Geboren wurde ich am 1. Januar 1932 in Insterburg in Ostpreußen als Sohn des Pfarrers Heinrich Wetterling und seiner Ehefrau Elisabeth geb. Pillardy. In Insterburg besuchte ich von 1938 bis 1942 die Volksschule und von 1942 bis 1944 das Gymnasium. Bei Kriegsende kamen meine Eltern mit mir nach Kassel. Dort besuchte ich bis zur Reifeprüfung im Jahre 1952 das Friedrichsgymnasium. Durch den Unterricht der Studienräte Brühne und Mentel hatte ich besonderes Interesse an der Mathematik gewonnen, die ich dann auch als Studienfach wählte, als ich im Sommersemester 1952 mein Studium in Frankfurt am Main begann. Nachdem ich dort hauptsächlich bei den Professoren Moufang und Behrens Vorlesungen gehört hatte, legte ich im Jahr 1954 die Vordiplomprüfung ab. Anschließend setzte ich das Studium in Hamburg fort. Angeregt durch die Vorlesungen von Professor Collatz beschäftigte ich mich besonders mit der Angewandten Mathematik. Im Frühjahr 1957 bestand ich die Diplomprüfung mit einer Arbeit über Eigenwertaufgaben bei gewöhnlichen Differentialgleichungen. Anschließend arbeitete ich am Institut für Angewandte Mathematik der Universität Hamburg zunächst an einem Forschungsauftrag über partielle Differentialgleichungen, dann als Assistent an der elektronischen Rechenanlage, die dem Institut von der Deutschen Forschungsgemeinschaft zur Verfügung gestellt wurde. Durch diese Arbeit erhielt ich die Anregung zu meiner Dissertation.

Herstellung
XEROGRAPHIE und ROTAPRINT -Offset
G.M.L.Wittenborn Söhne Hamburg 13